University of Zurich[UZH]

Department of Informatics

Designing Combinatorial Markets for Distributed Data

Dissertation submitted to the
Faculty of Business, Economics and Informatics
of the University of Zurich

to obtain the degree of
Doktor der Wissenschaften, Dr. sc.
(corresponds to Doctor of Science, PhD)

presented by
Dmitrii Moor
from Moscow, Russian Federation

approved in September 2019

at the request of
Prof. Sven Seuken, Ph.D.
Prof. Enrico Gerding, Ph.D.

The Faculty of Business, Economics and Informatics of the University of Zurich hereby authorizes the printing of this dissertation, without indicating an opinion of the views expressed in the work.

Zurich, September 18, 2019

Chairman of the Doctoral Board: Prof. Thomas Fritz, Ph.D.

Abstract

Similarly to how the Industrial Revolution was accompanied by rapid developments of markets for natural resources (such as oil and coal), the development of new markets for information goods is crucial to fully realize the benefits of the Digital Revolution. In particular, technological innovations of the last few decades resulted in the production of huge volumes of data by numerous businesses and individuals. Making such *distributed* data available and easily exchangeable between different parties leads to significant benefits for our society.

In this thesis, I consider the problem of designing a market for distributed data. I focus on the following specific features of this domain: (1) data providers have high costs for producing their databases and low costs for maintaining these databases; (2) users submitting their queries have combinatorial preferences over which databases are produced and get allocated; (3) the market exhibits a complex two-level production structure; (4) data providers and users can arrive to the market stochastically over time. These factors outline the *design space* for the market for distributed data that I consider in this thesis. I start with designing a number of mathematical models that tackle some particular aspects of this design space in isolation: from modeling combinatorial preferences of buyers to studying uncertainty regarding availability of goods. I proceed with a formal market design in a *static* setting (i.e., assuming that buyers and sellers arrive to the market at the same time). I argue for the need of regulation for such a market and propose an economic mechanism that can perform this regulation. I further study the most general domain with all four aforementioned features present, i.e, including the stochastic arrival of buyers and sellers to the market. To make the market design challenge tractable in this domain, I restrict the class of possible mechanisms to simple posted price mechanisms. The main challenge in this domain is to guarantee *dynamic incentive compatibility* (i.e., to prevent buyers and sellers from misreporting their costs and values as well as their arrival times in equilibrium). I further suggest a posted price mechanism that efficiently copes with this challenge, while guaranteeing zero expected average budget deficit.

Zusammenfassung

Ähnlich wie die industrielle Revolution von einer raschen Entwicklung der Märkte für natürliche Ressourcen (wie Öl und Kohle) begleitet wurde, ist die Entwicklung neuer Märkte für Informationsgüter von entscheidender Bedeutung, um die Vorteile der digitalen Revolution voll auszuschöpfen. Insbesondere die technologischen Innovationen der letzten Jahrzehnte führten dazu, dass zahlreiche Unternehmen und Einzelpersonen riesige Datenmengen produzierten. Die Bereitstellung solcher verteilten Daten und der einfache Austausch zwischen verschiedenen Parteien führen zu erheblichen Vorteilen für unsere Gesellschaft.

In dieser Arbeit beschäftige ich mich mit dem Problem der Gestaltung eines Marktes für verteilte Daten. Ich konzentriere mich auf die folgenden spezifischen Merkmale dieser Domäne: (1) Datenanbieter haben hohe Kosten für die Erstellung ihrer Datenbanken und niedrige Kosten für die Pflege dieser Datenbanken; (2) Benutzer, die ihre Abfragen einreichen, haben kombinatorische Präferenzen über die Datenbanken erstellt und zugewiesen werden. (3) Der Markt weist eine komplexe zweistufige Produktionsstruktur auf. (4) Datenanbieter und -nutzer können im Laufe der Zeit stochastisch dem Markt beitreten. Diese Faktoren definieren den Entwurfsraum für den Markt für verteilte Daten, den ich in dieser Arbeit betrachte. Ich beginne mit der Entwicklung einer Reihe mathematischer Modelle, die sich isoliert mit bestimmten Aspekten dieses Entwurfsraumes befassen: von der Modellierung kombinatorischer Präferenzen von Käufern bis zur Untersuchung der Unsicherheit in Bezug auf die Verfügbarkeit von Waren. Ich fahre mit einem formalen Market Design in einem statischen Umfeld fort (d. h. unter der Annahme, dass Käufer und Verkäufer gleichzeitig auf dem Markt eintreffen). Ich plädiere für die Notwendigkeit einer Regulierung für einen solchen Markt und schlage einen wirtschaftlichen Mechanismus vor, der diese Regulierung durchführen kann. Ich untersuche weiterhin den allgemeinsten Bereich mit allen vier oben genannten Merkmalen, d. h. einschliesslich der stochastischen Ankunft von Käufern und Verküufern auf dem Markt. Um die Herausforderung des Marktdesigns in diesem Bereich umsetzbar zu machen, beschränke ich die Klasse der möglichen Mechanismen auf einfache Mechanismen für veröffentlichte Preise. Die Hauptherausforderung in diesem Bereich besteht darin, eine dynamische Anreizkompatibilität zu gewährleisten (d. h. zu verhindern, dass Käufer und Verkäufer ihre Kosten und Werte sowie ihre Ankunftszeiten im Equilibrium falsch angeben). Ich schlage ferner einen Mechanismus für veröffentlichte Preise vor, der diese Herausforderung effizient bewältigt und gleichzeitig ein erwartetes durchschnittliches Haushaltsdefizit von Null garantiert.

Acknowledgements

To start with, I would like to thank my advisor, Prof. Dr. Sven Seuken who gave me the opportunity to work on this truly exciting project. I am very grateful for his constant support and countless discussions of the economic models, markets and algorithms. From Sven I have learned how to question assumptions, be diligent in research and how to structure and present my work. The research environment cultivated by Sven within the Computation and Economics Research Group was truly extraordinary and intellectually stimulating for the whole period of my doctoral studies.

I am also very thankful to Prof. Dr. Enrico Gerding who agreed to serve as an external reviewer for my thesis and who provided his valuable feedback for my Ph.D. proposal. This work would not be possible without the fruitful collaboration with Prof. Dr. Abraham Bernstein and Dr. Tobias Grubenmann, who enthusiastically shared with me lots of their technical expertise and were always eager to give me their very constructive feedback.

I wish to specially thank Prof. Dr. Ofer Shir and David Amid, with whom I had an opportunity to collaborate while working at IBM and finishing my studies in Moscow. Ofer and David are the two people who are the most responsible for me taking the challenge of doing my Ph.D. and I am grateful to them for their constant support and encouragements. Their passion towards science and their high standards for research have always inspired me. I would also like to thank Prof. Dr. Torsten Hoefler with whom I had a chance to work for one year at the Computer Science department of ETH Zurich. Torsten's dedication towards research and his inspiration about his field were truly contagious!

Special words of appreciation are dedicated to all participants and speakers of the 28th Jerusalem School in Economic Theory. Those two weeks at the Hebrew University had an enormous influence on my understanding of economics and of the world outlook generally, and were a game changing experience for my doctoral studies. I am also very grateful to IBM Research where I had an opportunity to spend three and a half months as an intern and in particular to Dr. Anika Schumann for letting this happen.

I would also like to express my gratitude to all my colleagues at the Computation and Economics Research Group: Gianluca Brero, Vitor Bosshard, Ludwig Dierks, Stefania Ionescu, Timo Mennle, Nils Olberg, Steffen Schuldenzucker, Mike Shann and Jakob Weissteiner. Your critical feedback and constructive suggestions for improvements helped me a lot to shape this work. I also want to thank Jiaoyan Chen, Juan Manuel Sanchez-Cartas, Jetlir Duraj and Timo Schneider for our numerous discussions about

the fundamentals of economics, mathematics and science in general. Finally, I wish to acknowledge the Swiss National Science Foundation for supporting this work.
On a personal level, I am very grateful to Gianluca Brero and Silvia Panero for their constant and unconditional support. It would be hard to wish for better friends than this couple! Surviving in one of the deadliest deserts on Earth with Gianluca and sharing the same office with him for three years made me believe that finishing my Ph.D. after all cannot be such a big issue.
I am particularly thankful to Angela Baciu who became the person that was always around in the hardest days. Switching between the roles of the inspiring teacher, caring friend and a strong leader, she taught me how to keep moving forward elegantly and confidently no matter what is happening around. Angela introduced me to a number of brilliant people who became my close friends and without whom this thesis would not be possible: Thilo Ladner, Dina Mayer, Cecile Python, Arseniy Klimovsky, Samia Nai, Inna Khriplovich, Valerie Erb and Naoko Yoshida. I wish to thank all of you for all the time we have spent together, our discussions, travels and for all your encouragements!
I am also specially grateful to Alexander Antukh, Arsenio Solovyev and my brother Andrei who have been the source of inspiration for me for already so many years. Thank you for giving me all your support and for making things so clear and simple to me when I persistently tried to turn flies into elephants. It is sometimes unbelievable how living on so different parts of the globe, we can keep such strong ties!
Most of all, I am infinitely thankful to my parents. Nadezhda and Aleksandr, I could have never done this without your endless love and unlimited support. You have taught me how to take risks, embrace any challenge and to never give up. Dad, it was you who always served to me as a role model and you keep astonishing me with your intelligence and high standards in everything you do. Mother, if there existed a superhero, then it must have been you. No matter how hard it was for you especially in your last years, you always shared your wisdom with me and encouraged me in anything I did. I fully devote this work to you.

To my parents, Nadezhda and Aleksandr.

Contents

1 Introduction and Overview of Results

1.1 Motivation[1]

1.1.1 Economics Meets Technology

The adoption of new ways of representing information has always gone hand in hand with an evolution in both technology and society. The invention of the printing press around 1440 A.C. by the German blacksmith Johannes Gutenberg turned the publication of books into an enterprise: It became much cheaper to produce multiple copies of a book, but at the same time, it required significant initial capital to build the printing press. These high *fixed costs* of production, as well as higher risks regarding future utilization, gave rise to a number of new market models for selling books from monopoly pricing, to the emergence of cooperative associations and syndicates, to subscription publishing, (Tucker et al., 2017). This symbiosis between technology and new economic markets stimulated a significant growth in the number of published books and periodical literature. It enabled European society to fully enjoy the numerous intellectual movements that would flourish in the upcoming Age of Enlightenment.
The synergies between technology and markets are not restricted solely to the example above. In fact, there are numerous cases in which these synergies benefited society by creating new ways of representing data. Consider, for example, the film, photography or sound recording industries. Rapid development of the underlying technologies became possible due to the highly profitable movie and entertainment markets that emerged in the second half of the last century.

1.1.2 Call for Data Markets

The technological progress of the second half of the last century led to the emergence of new markets for selling digital data. ‘‘Data is the new oil.” This topic has been repeatedly discussed by The Economist (2017b,a), The New York Times (2018) and a

[1]Parts of this section borrow from my previous papers Moor et al. (2019), Moor (2019), Moor et al. (2016).

number of other reputable publishers. It is not without reason that, in recent years, this message has been conveyed so persistently to the general public. Indeed, in the last decade, we have seen a spectacular growth in both the demand and the supply of digital data. The primary reason for this growth originates in the fast computerization that happens simultaneously in almost every aspect of our daily lives.

Indeed, the strong demand for digital data in recent years follows the explosive growth in scientific and industrial research in the area of machine learning. The vast majority of applications in this field heavily rely on enormous volumes of data. Data is typically used to train and validate machine learning models and to make certain inferences according to these models. However, it is not a purely theoretical interest that drives the active development in these areas. In fact, the last decade is marked by the spark of real-world applications that rely on machine learning techniques. The development of modern advertisement auctions that lie at the heart of Google's business model, as well as the invention of recommender systems heavily used by Amazon, are just the two well-known applications in this field.

The supply of digital data has exhibited a similarly explosive trend. According to a recent McKinsey report (2016), there are large amounts of data generated by industries adopting the Internet of Things (IoT) technologies. The IoT technologies inherently rely on hundreds to thousands of sensors that generate enormous flows of data. This data can be used to optimize business processes or perform predictive maintenance, among other tasks. The report states that while this data can be of high value for a number of different stakeholders, it is not currently used efficiently. This inefficiency arises from the fact that companies cannot easily buy or sell their data. This leads the authors to conclude that there is a strong need for markets for this kind of data.

Structured Data. The evidence coming from academia confirms the intuition above by promising great benefits to society if we publish more data in a *structured* way (i.e., as a database). This can allow machines to understand relationships between different pieces of data (see, for example, Bernstein et al. (2016)). Delegating this task to automatic query processing algorithms would significantly reduce the human effort required to analyze unstructured datasets. In the life sciences, for example, researchers submit queries that join data from databases provided by different companies. Each of these databases contains information on chemical compounds, disease data, biological function and biomarkers. Automatic aggregation and processing of this data leads to faster and more efficient drug discovery (see, for example, HCLS (2001)). Another illustrative example is IBM Watson, a large-scale question answering system that defeated human

champions on the popular TV show *Jeopardy*. This system heavily relies on querying structured data from distributed databases (Ferrucci et al., 2010).
Fortunately, many technologies that enable automatic aggregation and processing of structured data already exist. For example, the *Web of Data* (WoD), sometimes referred to as *Semantic Web* (W3C, 2014), is a stack of technologies that allows data providers to represent their data as a graph of abstract *concepts* connected by *relationships*. A user who submits a query to the WoD relies on a *query engine* to join several of those graphs produced by different *data providers*. In this case, the query engine does all the work of joining and processing the data and returns a direct answer to the user's query. Despite the apparent power of the WoD approach, the technology has not yet seen widespread adoption. One reason for this underutilization is economic in nature: most of the data produced for the WoD so far has either been subsidized by governments or produced at a loss (Buil-Aranda et al., 2013). This suggests that one of the most important factors preventing the widespread adoption of such technologies is the lack of financial incentives for data providers to publish their data in a structured way. Indeed, data providers may incur high costs for producing their databases, and they naturally hope to recoup these costs. However, advertising - the main source of income for many data publishers, as well as traditional search engines - does not work in this domain. This is because such structured and distributed data is processed by machines rather than by humans, and the machine can simply ignore any ad. Therefore, new sources of revenue for data providers are needed. This can be achieved by a market in which providers sell data to users and the trade is mediated by a market platform. In this thesis, I conduct research into how to find the best design for such a marketplace.

1.2 Challenges Designing Data Marketplaces

Despite the existence of numerous companies that sell their proprietary databases, there is still no clear understanding of what is the best design of a marketplace for distributed databases. Indeed, data published as a database can be easily linked against databases of other data providers. This potentially increases the value of this database for users as now they can submit complex queries against complementary databases. Therefore, to achieve additional gains from the trade, data providers should be interested in linking their databases among each other. However, we do not see this happening on large scale. To understand how to capture the welfare gains arising from complementarities of different databases, I study a number of economic and technical aspects of this domain that are the most relevant for designing a practical marketplace for distributed data. In

this thesis, I focus on the following aspects of the design space.

High fixed costs, low marginal costs. Similarly to other markets for digital goods, data providers in markets for distributed data typically have high *fixed* costs for producing the first instance of their databases and very low *marginal* costs for maintaining their databases. In this case, the fixed costs correspond to the costs of structuring the data, linking it against databases of other data providers, setting up the database, etc. The marginal costs include the electricity costs of running the database, answering queries, etc. As the willingness of users to pay for answers for their queries is typically much lower than the fixed costs of production of databases, these costs can be recouped only by collecting enough money from some (potentially large) number of users.

Combinatorial preferences of users. One peculiarity of the domain of data markets is that the databases produced by different data providers can be complementary for buyers. Consider the following example. There is a buyer who wishes to drive to New York to have a dinner.[2] Assume that there are two databases available to the buyer. The first database contains the list of addresses of different restaurants in New York, while the second contains the number of available parking slots next to each restaurant. The buyer who can query only the first database may be willing to pay some positive amount of money for it. Indeed, having received the list with addresses of all the restaurants in New York, the buyer can simply check each of the restaurants until he finds one with an available parking slot. However, if the buyer can join both databases, he gets a precise answer for his original question of where to drive to have dinner. Thus, his value for accessing the two complementary databases must be higher. Designing a market that allows to achieve additional gains by exploiting this combinatorial aspect of buyers' preferences is, therefore, a crucial design decision for the domain of distributed data.

Complex production structure. In the domain of distributed databases, it makes sense to consider the market as a production economy rather than as an exchange economy. Interestingly, production in these markets happens on two distinct levels.
On the first level, it is data providers that produce their databases. The second level of production in data markets corresponds to the production of answers for buyers' queries out of the databases of potentially different data providers. It follows that the goods produced by data providers (i.e., databases) differ from the goods consumed by buyers (i.e., answers to their queries). This differentiates the domain of data markets from many

[2]To simplify the language, we use "he" for the buyer (user) and "she" for the seller (data provider).

other domains in which using a simple auction format for redistribution of the goods may be sufficient.

Dynamic arrival of buyers and sellers. While in many combinatorial markets the dynamics may not play a critical role, data markets are inherently dynamic.[3] This means that both buyers and sellers in these markets arrive regularly and can strategically delay their arrivals if they expect to be better off by doing so. Due to the combinatorial nature of preferences of buyers such delays can have a dramatic effect on the operation of the market. Indeed, the late arrivals of sellers may result in a very low surplus reached by the buyers who arrive earlier and thus, can access only very few databases.

Each of the aforementioned aspects is accompanied by a number of computational and economic market design challenges. In this thesis, I suggest a number of market models that efficiently cope with those challenges.

1.3 Goal and Research Questions of the Thesis

This thesis focuses on the development of mathematical (game-theoretical) models that describe the basic economics of markets for distributed data. The ultimate goal is to design the economic mechanisms that govern trade in such markets. These mechanisms should build the foundation for more practical market design solutions in this domain in the future. While the overall challenge of practical market design for distributed data may sound ambitious, the underlying economic mechanisms must provide a good enough *approximation* for these practical marketplaces. Thus, they must have provable theoretical guarantees on their optimality or be supported by extensive computational analysis if a full theoretical examination is infeasible. From the computer science perspective, the proposed mechanisms should be either efficiently computable or provide appropriate approximations. Apart of that, the field of computer science and particularly of distributed databases provides us with the most important technological constraints for the market design challenge.

As is common in economic mechanism and market design, I adopt the perspective of a market designer who aims at building highly *efficient* markets (i.e., markets with high levels of surplus reached by the participants) whenever this is possible. To ensure that the proposed market is practical and to be able to formally reason about its economic

[3]For example, combinatorial spectrum auctions typically happen once in several years (Cramton, 2013). Within this time frame the technology can change dramatically making it impractical for the bidders to misreport their bids based on the expected outcome of one of the future auctions.

properties, this must be subject to a number of additional constraints - namely, one should maintain *incentive compatibility*, which ensures robustness of the market to strategic manipulations by both users and data providers. A second property, *individual rationality*, guarantees that the market participants do not end up with negative utilities if they participate in the market. In particular, it means that the data providers can recoup the fixed costs they incurred when producing their databases, and it also ensures that users obtain a non-negative *surplus* when participating in the market. Finally, *budget balance* is the property that allows the market to operate without external subsidies. Unfortunately, there does not exist a mechanism for satisfying all these desiderata in an arbitrary domain. This follows, from the seminal paper by Myerson and Satterthwaite (1983). Instead, the design of the market must be tailored to the particular nature of *distributed data*. This raises the following three research questions:

Question 1. What are the technological constraints of the domain of distributed data and what are the economic features of such a market?

Question 2. How can we formally design a *static* mechanism that can cope with the constraints arising from Question 1?

Question 3. How can we model the *dynamic* arrival of buyers and sellers to the market and design an appropriate mechanism for mediating the trade in a dynamic setting?

1.4 Publications Contained in this Thesis

This thesis consists of four papers that answer the three research questions presented in Section 1.3. In this section, I restate the research questions and provide a list of papers that address the respective research questions.

Question 1. What are the technological constraints of the domain of distributed data and what are the economic features of such a market?

Publications:

- **D. Moor**, T. Grubenmann, S. Seuken, A. Bernstein (2015). A Double Auction for Querying the Web of Data. In *Proceedings of the Third Conference on Auctions, Market Mechanisms and Their Applications*, AMMA, Chicago, IL, USA.

- **D. Moor**, S. Seuken, T. Grubenmann, A. Bernstein (2016). Core-Selecting Payment Rules for Combinatorial Auctions with Uncertain Availability of Goods. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence*, IJCAI, New York, NY, USA.

Question 2. How can we formally design a *static* mechanism that can cope with the constraints arising from Question 1?

Publications:

- **D. Moor**, S. Seuken, T. Grubenmann, A. Bernstein (2019). The Design of a Combinatorial Data Market. *Working paper*.

Question 3. How can we model the *dynamic* arrival of buyers and sellers to the market and how to design an appropriate mechanism for mediating the trade in the dynamic setting?

- **D. Moor** (2019). Data Markets with Dynamic Arrival of Buyers and Sellers. In *Proceedings of the 14th Workshop on Economics of Networks, Systems and Computation*, NetEcon, Phoenix, AZ, USA.

Additional work. The following papers I co-authored shed some more light on the technical aspects of the domain of distributed data (Question 1). However, I do not provide the full content of these papers in my thesis as I am not the first author of those.

- T. Grubenmann, A. Bernstein, **D. Moor** and S. Seuken (2017). Challenges of source selection in the WoD. In *Proceedings of the 16th International Semantic Web Conference*, ISWC, Vienna, Austria.

- T. Grubenmann, A. Bernstein, **D. Moor**, S. Seuken (2018). Financing the Web of Data with Delayed-Answer Auctions. In *The Web Conference*, WWW, Lyon, France.

- T. Grubenmann, A. Bernstein, **D. Moor**, S. Seuken. FedMark: A Marketplace for Federated Data on the Web. *Working paper*.

1.5 Related Work

1.5.1 Existing Business Models

I here mention a number of companies that operate marketplaces for data in certain business areas. I then briefly summarize their respective business models, the kind of data that they provide and the typical data providers and consumers of this data.

Bloomberg. Bloomberg L.P. is a privately held company that makes most of its profits operating as an aggregator of financial data and analytics. The company sells access to the *Bloomberg Terminal* to banks, hedge funds, brokers and other financial institutions. This terminal allows easy access to and display of financial data and news, and it provides a number of analytical tools. The platform bundles together numerous databases from both free and proprietary sources. Free databases typically contain data that is *common knowledge*, such as simple stock prices. A typical provider of the proprietary data is the New York Stock Exchange (NYS) and Comscore. Bloomberg terminals are typically leased for one or two years. The subscription costs between 20,000 USD and 25,000 USD per year (Greg, 2011). According to Quartz (2013), the company suggests discounts depending on the number of terminals acquired.

Thomson Reuters. Thomson Reuters is widely considered to be Bloomberg's main competitor in the market for trading financial data. In addition to financial data, the company also provides software and numerous databases with legal data used by lawyers, accountants and journalists. According to the annual report of Thomson-Reuters (2015), the largest share of its operating revenue is derived from subscription leases. The company also suggests discounts depending on the total contract price: the higher the price, the larger the discount.

LexisNexis. LexisNexis provides access to a number of legal databases. Typical buyers include law firms, banking and financial services companies and state and local governments. Data providers for LexisNexis include other companies - for example, Thomson Reuters. A number of databases offered by LexisNexis are also produced by the company itself. In contrast to Bloomberg and Thomson Reuters, LexisNexis suggests transactional pricing in addition to flat rate pricing. With transactional pricing, customers can pay either per search or per hour. Similarly to Blomberg and Thomson Reuters, LexisNexis offers discounts to certain customers (e.g., governments). Under the flat rate pricing,

Lexis Nexis suggests two subscription tariffs with some degree of content restriction (LexisNexis, 2019).

Microsoft Azure Data Marketplace. A practical data marketplace was attempted by Microsoft with its Microsoft Azure Data Marketplace platform, but it ceased operations in 2016 due to "lack of sustained customer interest" (Ramel, 2016). This lack of interest, however, does not necessarily suggest a lack of demand; an inadequate business model could be to blame. Evidence for this is the fact that companies like Thomson Reuters, LexisNexis and Bloomberg still make large profits by selling access to their proprietary databases. Still, Microsoft Azure Data Marketplace was a remarkable example of a data market, as it allows any data provider to deliver its database to the platform (i.e., the data is not restricted to a particular domain).

Other. There are a number of smaller companies that compete with Bloomberg and Thomson Reuters in the market for financial data. FactSet is just one example. Similarly, in the domain of legal data, Wolters Kluwer is a competitor of LexisNexis. There are also a number of other domains apart from those discussed above, with numerous companies selling their data. For example, Elsevier is a well known publisher of scientific and technical databases. A good overview of different existing data vendors and data market places can be found in Schomm et al. (2013).

1.5.2 Related Literature

There are numerous papers focusing on designing markets for digital goods as opposed to non-digital goods. Varian (1995, 1997) discusses the problem of competition between producers of digital goods in an environment with high *sunk* and low marginal costs. In Varian (2000), the problem of buying and sharing of information goods is analyzed. Goldberg et al. (2001), Goldberg and Hartline (2001, 2003) propose an auction for selling goods in unlimited supply (such as music or video). Bakos and Brynjolfsson (1999) showed that in domains with high uncertainty about the valuations of users for information goods, bundling allows the seller to better predict the value of buyers for the bundle and, thus, to extract more revenue. All aforementioned approaches discuss a general *digital good* such as music or video files. Unfortunately, none of these approaches allow to model the combinatorial structure of users' preferences that is specific for the domain of distributed databases (see Section 1.2).
Generally speaking, it is not straightforward that buyers' preferences in markets for distributed data can be described by a utility function. However, there are a number of

cases in which deriving the utility of a buyer is possible. For example, a recent paper by Bergemann et al. (2018) presents a model in which heterogeneous buyers need to take an action in an uncertain world. The buyers do not know in which state the world is, but they have private prior beliefs about this state. As the payoff of the buyers' decisions depend on the state of the world, they can acquire some supplementary data to update their beliefs. Thus, the value of a buyer for the data can be defined as the gain in utility that the buyer expects to get from the refined signal about the state of the world. Consequently, the utility can be described as the value of the buyer for this information, net the payment that the buyer has to make to acquire the data.

A different model for the value of information was proposed by Admati and Pfleiderer (1986). In their model, a number of homogeneous traders acquire information regarding an asset that is traded in a speculative market. The information received by traders is reflected in the price of the asset and, thus, the value of the information decreases as more participants in the market gain access to the information. This value model is conceptually different from the one proposed by Bergemann et al. (2018) in how the value of information depends on the number of buyers to whom the information is revealed. While this thesis does not contribute directly to this line of research, it is build upon the idea that the buyers in data markets are able to estimate their values for data.

There are numerous papers discussing how to sell structured data (such as a single database). For example, Mehta et al. (2019) propose a utility model and an optimal mechanism for selling a single dataset that consists of multiple rows and columns. The authors have shown that despite the heterogeneity of the buyers, the optimal mechanism is a fairly simple ‘‘price-quantity" schedule. In Koutris et al. (2015), Balazinska et al. (2013), Deep and Koutris (2016) the authors design a query-based pricing. As different queries can lead to potentially different result sets, the number of different goods, in this case, coincides with the number of different queries that the user can submit. Obviously, this number can be large which opens a possibility of *arbitrage* in this domain (i.e., the possibility to produce a similar *view* of a database for a smaller payment), which these authors pay special attention. Consequently, designing a query-based pricing rule for such domains needs some special care. None of these works, however, models the fact that the data can be produced by different data providers. In the domain of distributed databases, this is a crucial feature and my thesis takes this into account.

This idea of combinatorial preferences in markets for distributed data is further emphasized by Agarwal et al. (2019) who suggested a possible solution for a marketplace that sells data for machine learning tasks. In their paper, the authors proposed a design of a marketplace for a setting when data providers have already incurred high sunk costs for

producing their data and would naturally like to minimize their regret. In contrast to that, in my work, I aim at designing a marketplace that would incentivize sellers who have not yet produced their data to do so. In particularly, I want to guarantee that these sellers always get fully compensated for producing their databases.[4]

1.6 Summary of Contributions

In this section, I briefly summarize the four papers mentioned in Section 1.4, which together constitute the main contribution of this thesis.

1.6.1 A Double Auction for Querying the Web of Data

The paper presents a possible market design solution for the domain of the Web of Data (WoD) in a setting in which sellers have considerable marginal costs for answering queries. The solution is based on using a combinatorial double auction to allocate different sets of sellers to answer queries submitted by buyers. In this setting, each buyer submits his bid to the market platform. The bid consists of the query and a willingness to pay for the query. Naturally, the buyer can misreport his willingness to pay. However, I assume that the buyer cannot misreport his query. Similarly, data providers submit their cost estimates for answering the queries, as well as statistics regarding the data at their disposal. Data providers can lie about their costs (i.e., to submit lower or higher costs than the true costs). However, the market platform can validate the statistics and, thus, the data providers cannot manipulate it.

The main challenge discussed in the paper arises from the uncertainty of the actual (realized) results of query execution. Indeed, when the buyer submits his query, he is not aware of the actual result set that he will ultimately receive. Thus, upon receiving the result set of a lower value than expected, the individual rationality constraint for the buyer may be violated. To eliminate this, we must correct the payments computed by the market platform *ex-post*. In this paper, I present a number of possible payment correction rules.

Important here is that the double auction discussed in the paper aims at maximizing social welfare in the two-sided market. To achieve this goal however, we must sacrifice strategyproofness (this is due to the result of Myerson and Satterthwaite (1983)). Therefore, to compare the different payment correction rules I perform a computational

[4]These two different market design objectives result in different amounts of information that the market designer needs to know to compute allocation and payments. In particular, the design of Agarwal et al. (2019) does not rely on prior knowledge of distributions of sellers' costs while my design does.

Bayes-Nash equilibrium analysis of the proposed market. The analysis showed that the proposed payment correction rules perform relatively well (with respect to the efficiency reached in the market) comparing to the *Threshold* rule that is commonly used in a double auction setting without uncertainty (Parkes et al., 2001). This suggests that implementing an auction-based market for the WoD may be a viable solution. The main limitation of the model, however, is that it assumes high marginal costs of answering queries and ignores the fixed costs that the data providers incur for producing their databases.

1.6.2 Core-Selecting Payment Rules for Combinatorial Auctions with Uncertain Availability of Goods

In this paper, I present the design of a combinatorial auction for a domain in which goods can become unavailable. While this problem may seem to be unrelated to the problem of designing the market for distributed data, it has numerous similarities. First, in both domains, buyers' preferences are combinatorial. Second, in the two domains the goods that the buyer expects to receive may differ from the goods he actually receives (i.e., similarly to Moor et al. (2015), the buyer can receive a different result set than he expects).

In this case, I follow the general approach of Porter et al. (2008) and assume that each good can become unavailable with a certain probability. I further suggest the design of a core-selecting combinatorial auction that maximizes the expected social welfare (with respect to the reported bids) and computes the payments *ex-post*. This means that the payments computed by such an auction are *contingent* on the actual realization of availabilities of different goods. One of the main results of this paper is that generally, there does not exist a mechanism that is ex-post core-selecting and budget balanced. This follows from the fact that the ex-post core can be empty. However, in practice we can circumvent this issue by computing the payments from the ex-post core whenever the core is non-empty and using some reference payments (such as VCG) in all other cases.

As core-selecting auctions are generally not strategyproof, I analyze the performance of the proposed mechanism via a computational Bayes-Nash equilibrium analysis. The analysis suggests that the proposed payment rules perform much better than a naive application of a standard combinatorial auction to the domain with uncertainty. In addition to that, the rate of violations of the individual rationality constraint gets smaller for the ex-post computation of payments comparing to the ex-ante computation. The

suggested approach, however, has a number of practical limitations. The main limitation is the high computational complexity in large settings. Observe also that this setting corresponds to a one-sided market and, thus, assumes a single auctioneer (seller).

1.6.3 The Design of a Combinatorial Data Market

This paper suggests a practical market design solution for a static case in which buyers and sellers arrive to the market at the same time. In this setting, I assume that each seller can produce at most one database and has a high fixed cost for producing it. A buyer can submit his query to the query engine that resides on the market platform's side. An answer for the buyer's query constitutes the *result set* which typically consists of multiple rows. In this case, the value of a buyer for the result set depends on how many rows does the result set contain as well as on the databases that are used to produce the result set.

The design exploits the idea of the market platform playing the role of a *regulator* that aims at achieving high levels of surplus reached by buyers while guaranteeing that the fixed costs of allocated sellers are recouped. In this setting, the market platform computes the allocation of sellers as well as the payments that should be paid to the sellers. This is done using a Buyer Optimal Reverse Auction (BORA) that is designed to optimize the surplus of buyers. The market platform also determines the posted price (per row of the result set) that is exposed to buyers. The main challenge in this setting is keeping both parts of the market balanced (i.e., to ensure that the total amount of money collected from the buyers is sufficient to compensate the allocated sellers). This is done by designing an appropriate fixed point iteration procedure that gradually updates the posted price and recomputes the outcome of the BORA auction as long as the budget deficit is not equal to zero.

To validate the model I provide an extensive simulation study of the proposed mechanism in small- and medium-sized domains. The results of this study confirm the high efficiency achieved by the proposed model. I also demonstrate that in such a market the sellers have a strong incentive to "innovate", i.e., to produce unique databases rather than to compete with each other for producing databases out of *common knowledge* data. Development of large-scale simulations and testing the market under real-world conditions is one possible future extension of this work.

1.6.4 Data Markets with Dynamic Arrival of Buyers and Sellers

The focus of this paper is on modeling the dynamic arrival of buyers and sellers to the data market. Similarly to Moor et al. (2019), the market platform in this model plays the role of a regulator that aims to optimize the surplus of buyers while guaranteeing that the fixed costs of the allocated sellers are recouped. The operation of the market platform is modeled by a Markov process, with states corresponding to different numbers of allocated databases. The transition probabilities are defined by posted prices that the market platform exposes to data providers as well as by the arrival rate of the new data providers.

This paper introduces the notion of *dynamic incentive compatibility* that prevents late arrivals of agents. To guarantee dynamic incentive compatibility, the posted prices exposed to buyers must differ from those computed for sellers. In the paper, I argue that optimal posted prices computed by the market platform and exposed to sellers can only decrease, which guarantees dynamic incentive compatibility for sellers. I also demonstrate that the posted prices exposed to buyers constitute a martingale process. This suggests that the expected price faced by the buyer tomorrow is equal to the posted price observed by the buyer today. This allows to achieve dynamic incentive compatibility for buyers. This paper also shows that dynamic incentive compatibility is generally not compatible with budget balancedness. Fortunately, as I show, the average expected budget deficit goes towards zero as the number of sellers in the market increases.

1.7 Conclusion and Future Work

Today, a thorough analysis of the data at hand is essential for any efficient policy and decision making. Ensuring that the data is readily available to decision makers is, therefore, a challenge with enormous potential gain. The goal of this thesis is to build the foundation for a practical market design that would incentivize data providers to publish and link their databases against the databases of other data providers while making it attractive for users of the data to participate in the market. Such a design must enable achieving additional welfare gains by allowing buyers to join databases across different domains. In this domain, budget balancedness is an important feature that makes such a market sustainable.

I performed an extensive exploration of the design space and proposed several economic mechanisms that cope with a number of market design challenges arising in the domain of distributed data. I have proposed the design of an optimal static mechanism for such

a market. This mechanism aims to maximize buyers' surplus while guaranteeing that allocated sellers are compensated for the fixed costs they incurred. It also guarantees budget balancedness and individual rationality.
Further, I studied the dynamics of such markets. In this realm, I identified the main challenges arising from the dynamic arrival of buyers and sellers on the market (e.g., dynamic incentive compatibility) and suggested a mechanism that efficiently copes with those challenges. These findings suggest that markets for distributed data, while extremely complicated, are still viable and can function highly efficiently under appropriately designed regulations.

1.7.1 Limitations

There are a number of market design challenges that have not been addressed in this thesis. For example, none of our current models allow for the measurement of the impact of privacy on the operation of the market. Some aspects of the privacy issues inherent to data markets have been studied by Jones and Tonetti (2018). In their work, the authors suggested a model in which multiple firms can trade personal data of their users. However, there has not been significant progress in understanding what this would suggest for practical market design on the level of data marketplaces.
This thesis also does not examine how different terms of use or qualities of data influence the efficiency of the proposed markets. While I also do not consider expiring data, this can be modeled within my Moor (2019) paper by adding backwards transition probabilities into the Markov chain of the market platform. However, I leave this direction for future work.

1.7.2 Future Work

Designing a centralized market for distributed data is a challenge. Once we have built the basic static and dynamic models of such a market, a natural next step would be to relax certain assumptions and to study the robustness of our models with respect to these relaxations. In particular, one could think of modeling more complex preferences of buyers by making these preferences dependent on the total number of buyers who can access a certain dataset (similarly to Admati and Pfleiderer (1986)). Modeling the implications of privacy and different qualities of data on the efficiency of the proposed market mechanisms would be another future direction. Notice also, that in practice, different databases can have different terms of use. Modeling this aspect of data markets may lead to further combinatorial challenges in query execution and potentially in pricing.

Thus, it constitutes another possible line of the future research.
An alternative direction for the future work would be to study the existing decentralized markets for data. Indeed, in our daily lives, we see a number of companies selling their data in a decentralized manner (see Section 1.5.1). This decentralized market evolved via a number of merges or acquisitions of smaller companies by companies like Bloomberg and Thomson Reuters. As a result, the current decentralized market for data exhibits a number of data aggregators in specific business areas. Understanding how such a market structure evolved, what is the optimal number and size of such aggregators and what prevents these aggregators from further merging may shed some light on the most critical obstacles preventing data markets to achieve full efficiency.

Bibliography

NYSE Exchange Proprietary Market Data. https://www.nyse.com/market-data.

Anat R Admati and Paul Pfleiderer. A monopolistic market for information. *Journal of Economic Theory*, 39(2):400 -- 438, 1986.

Anish Agarwal, Munther Dahleh, and Tuhin Sarkar. A marketplace for data: An algorithmic solution. Technical report, Massachusetts Institute of Technology, Working Paper, 2019.

Yannis Bakos and Erik Brynjolfsson. Bundling information goods: Pricing, profits, and efficiency. *Management Science*, 45(12):1613--1630, 1999.

Magdalena Balazinska, Bill Howe, Paraschos Koutris, Dan Suciu, and Prasang Upadhyaya. A discussion on pricing relational data. *In Search of Elegance in the Theory and Practice of Computation*, 8000:167--173, 2013.

Dirk Bergemann, Alessandro Bonatti, and Alex Smolin. The design and price of information. *American Economic Review*, 108(1):1--48, January 2018.

Abraham Bernstein, James Hendler, and Natalya Noy. A new look at the semantic web. *Commun. ACM*, 59(9):35--37, August 2016.

Carlos Buil-Aranda, Aidan Hogan, Jürgen Umbrich, and Pierre-Yves Vandenbussche. Sparql web-querying infrastructure: Ready for action? In *Proceedings of the 12th International Semantic Web Conference - Part II*, ISWC '13, pages 277--293, New York, NY, USA, 2013.

Comscore. Comscore. https://www.comscore.com/.

P. Cramton. Spectrum auction design. *Review of Industrial Organization*, 42(2):030--190, March 2013.

Shaleen Deep and Paraschos Koutris. The design of arbitrage-free data pricing schemes. Technical report, University of Wisconsin-Madison, Working Paper, 2016.

David Ferrucci, Eric Brown, Jennifer Chu-Carroll, James Fan, David Gondek, Aditya A. Kalyanpur, Adam Lally, J. William Murdock, Eric Nyberg, John Prager, Nico Schlaefer, and Chris Welty. Building Watson: An overview of the DeepQA project. *AI Magazine*, 31(3):59--79, 2010.

Andrew Goldberg and Jason Hartline. Competitive auctions for multiple digital goods. *Springer Berlin Heidelberg, Berlin, Heidelberg*, pages 416--427, 2001.

Andrew V. Goldberg and Jason D. Hartline. Envy-free auctions for digital goods. In *Proceedings of the 4th ACM Conference on Electronic Commerce*, EC '03, pages 29--35, New York, NY, USA, 2003.

Andrew V. Goldberg, Jason D. Hartline, and Andrew Wright. Competitive auctions and digital goods. In *Proceedings of the Twelfth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '01, pages 735--744, Philadelphia, PA, USA, 2001.

MacSweeney Greg. Inside the bloomberg machine. WallStreet and Technology, http://www.wallstreetandtech.com/trading-technology/inside-the-bloomberg-machine/d/d-id/1264634?, March 2011.

HCLS. Semantic Web Health Care and Life Sciences (HCLS) Interest Group. https://www.w3.org/2001/sw/hcls/, 2001.

Charles Jones and Christopher Tonetti. Nonrivalry and the Economics of Data. Technical report, 2018.

Paraschos Koutris, Prasang Upadhyaya, Magdalena Balazinska, Bill Howe, and Dan Suciu. Query-based data pricing. In *Journal of the ACM (JACM)*, volume 62, October 2015.

LexisNexis. Pricing Plans. https://www.lexisnexis.com/en-us/terms/21/pricing.page, 2019.

McKinsey report. Creating a successful Internet of Things Data Marketplace. http://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/creating-a-successful-internet-of-things-data-marketplace?cid=soc-web, October 2016.

Sameer Mehta, Milind Dawande, Ganesh Janakiraman, and Vijay Mookerjee. How to sell a dataset? pricing policies for data monetization. Technical report, The University of Texas at Dallas, Naveen Jindal School of Management, Working Paper, 2019.

Microsoft Azure Data Marketplace. https://datamarket.azure.com/home.

Dmitry Moor. Data markets with dynamic arrival of buyers and sellers. Technical report, Faculty of Business, Economics and Informatics, University of Zurich, Working Paper, 2019.

Dmitry Moor, Tobias Grubenmann, Sven Seuken, and Abraham Bernstein. A Double Auction for Querying the Web of Data. In *Proceedings of the Third Conference on Auctions, Market Mechanisms and Their Applications (AMMA)*, Chicago, USA, August 2015.

Dmitry Moor, Sven Seuken, Tobias Grubenmann, and Abraham Bernstein. Core-selecting payment rules for combinatorial auctions with uncertain availability of goods. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pages 424 -- 432, New York, NY, USA, 2016.

Dmitry Moor, Sven Seuken, Tobias Grubenmann, and Abraham Bernstein. The design of a combinatorial data market. Technical report, Faculty of Business, Economics and Informatics, University of Zurich, Working Paper, 2019.

Roger B. Myerson and Mark A. Satterthwaite. Efficient mechanisms for bilateral trading. *Journal of Economic Theory*, 29(2):265--281, April 1983.

David C. Parkes, Jayant Kalagnanam, and Marta Eso. Achieving budget-balance with vickrey-based payment schemes in exchanges. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'01, pages 1161--1168, San Francisco, CA, USA, 2001.

R. Porter, A. Ronen, Y. Shoham, and M. Tennenholtz. Fault tolerant mechanism design. *Artificial Intelligence*, 172(15):1783 -- 1799, 2008.

Quartz. This is how much a Bloomberg terminal costs. https://qz.com/84961/this-is-how-much-a-bloomberg-terminal-costs/, May 2013.

Ramel. Microsoft closing azure datamarket. Application Development Trends Magazine, https://adtmag.com/articles/2016/11/18/azure-datamarket-shutdown.aspx, November 2016.

Fabian Schomm, Florian Stahl, and Gottfried Vossen. Marketplaces for data: An initial survey. *ACM SIGMOD*, 42(1):15--26, 2013.

The Economist. Fuel of the future. Data is giving rise to a new economy. https://www.economist.com/news/briefing/21721634-how-it-shaping-up-data-giving-rise-new-economy, May 2017a.

The Economist. The worlds most valuable resource is no longer oil, but data. https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data, May 2017b.

The New York Times. As Facebook Raised a Privacy Wall, It Carved an Opening for Tech Giants. https://www.nytimes.com/2018/12/18/technology/facebook-privacy.html, December 2018.

Thomson-Reuters. Thomson reuters annual report. Technical report, 2015.

David H. Tucker, George Unwin, and Philip Soundy Unwin. History of publishing. *Encyclopedia Britannica*, February 2017. URL `https://www.britannica.com/topic/publishing`.

Hal R. Varian. Pricing information goods. Technical report, University of Michigan, 1995.

Hal R. Varian. Versioning information goods. Technical report, University of California, 1997.

Hal R. Varian. Buying, sharing and renting information goods. *The Journal of Industrial Economics*, 48(4):473--488, 2000.

W3C. Linked Data. https://www.w3.org/standards/semanticweb/data, 2014.

2 A Double Auction for Querying the Web of Data

> A VERY common form of motion of mechanical systems is what are called *small oscillations* of a system about a position of stable equilibrium. We shall consider first of all the simplest case, that of a system with only one degree of freedom.
>
> L.D. Landay and E.M. Lifshitz, Volume 1 of Course of Theoretical Physics

The content of this chapter has previously appeared in:[1]

> Moor, D. and Grubenmann, T. and Seuken, S. and Bernstein A. (2015). A Double Auction for Querying the Web of Data. In *Proceedings of the Third Conference on Auctions, Market Mechanisms and Their Applications*, AMMA, Chicago, IL, USA.

[1] This excludes the addendum added in the Section 2.1 of the current thesis.

A Double Auction for Querying the Web of Data

Dmitry Moor Tobias Grubenmann Sven Seuken Abraham Bernstein
Department of Informatics
University of Zurich

July 23, 2015

Abstract

Currently, the Web of Data (WoD) suffers from a lack of financial incentives for data providers. In this paper, we address this issue, by proposing a double auction to efficiently allocate answers (from data providers) to queries in the WoD. However, our domain exhibits a number of complicating features. Most importantly, before executing a particular query, the market mechanism only has estimates regarding what result can be expected. Thus, in contrast to other domains, the allocation rule as well as the pricing rule of the auction must operate based on *value* and *cost estimates*. New challenges arise from this setting; in particular the auction's *participation* constraint can no longer be guaranteed to be satisfied. We propose three payment correction rules to address this issue, and compare the efficiency of the resulting payment rules via a *computational Bayes-Nash equilibrium* analysis.

1 Introduction

The *Web of Data (WoD)* is to computers what the traditional Web is to humans. The goal is to expose data in a semantic format such that machines can easily find the information they are looking for. This semantic format allows for easy data integration and hence content from different sources can be queried in a federated fashion without the need to agree on a common scheme – something that is neither possible in the traditional Web nor in a traditional database setting. If implemented properly, a *semantic search* would lead to the desired results much faster than using a traditional search engine. However, a big practical challenge is the adoption of WoD technologies (Antoniou and van Harmelen, 2004). In particular, the majority of data that exists somewhere in some form is simply not made available in a WoD format.

The primary reason for this is a lack of incentives for the data providers: on the WoD, datasets are usually queried by algorithms rather than viewed by people. Thus, *advertising*, the main source of income for search engines on the traditional web, does not work in this environment where machines process the data and automatically filter out unwanted information or advertisement.

It is of course possible to directly charge users to access (and search through) high-quality data. For example, *Bloomberg*, *LexisNexis*, and *Thomson Reuters* charge customers high fees for accessing their data using a subscription-based model. More recently, marketplaces like the Azure DataMarketplace[1] has enabled different publishers to sell their data with different subscriptions based on the number of transactions per month. However, none of these companies provide their data in a way such that they can be queried in a federated fashion. They do not provide the means to join multiple data sets, thereby forgoing the complementarities the WoD would enable.

[1] https://datamarket.azure.com/home

This is a serious drawback, because customers are often interested in a specific *combination* of data from different providers that are joined in a certain way.

1.1 Market Incentives for Supplying Data

To create new incentives for providing data in a semantic format, we propose to use a market for querying the WoD. Specifically, we suggest using a double auction to elicit buyers' values for getting an answer to their queries as well as sellers' costs for supplying their data. Our goal in designing the auction is to elicit the (true) values and costs of the market participants such that we can make (approximately) efficient allocations, while satisfying the participants' participation constraints and keeping the market *budget-balanced*. Interestingly, in our domain, satisfying these design goals is highly non-trivial.

One argument against the marketplace's viability is that caching/duplication of data and Bertrand competition will drive prices down to zero, making it uninteresting for sellers to even enter the market. In practice, however, many data providers can sell information due to 1) frequent content changes (e.g., airplane schedules, today's movies, stock prices), 2) ever-increasing content (e.g., court cases, patent information or pharma information), or 3) licensing restrictions (e.g., restricting re-sale and sometimes even re-use of movies, music, or financial exchange information). Thus, there are many data sources for which our proposed market would be viable.

1.2 Market Design Challenges

The first complicating feature of our domain is that the sellers are selling their *data* instead of just *computational resources*. This means that a buyer who submits a query to the market does not know a priori what he will get in return; i.e., he is effectively "buying something of uncertain value." Concretely, the *goods* for sale in the auction are not fully specified upfront. This requires the buyer to specify his value function, which is a function that defines his value for different outcomes of the auction, for all possible result sets that the market could return.

A second, and even more central challenge in our domain, is the fact that the sellers can only provide *statistics* about the data they are selling (i.e., the goods in the auction); but based on these statistics, the market mechanism can only compute rough *estimates* regarding what result a specific allocation will produce. In particular, the exact value of an allocation will only be known after the sellers have produced the result. In consequence, the *allocation rule* and the *pricing rule* of the market mechanism must operate based on *value and cost estimates* – a distinguishing feature of our domain. We will demonstrate that this feature implies that for a two-sided market mechanism which is otherwise guaranteed to satisfy budget-balance and participation (Parkes, Kalagnanam and Eso, 2001), the participation constraint may now become infeasible to be satisfied, and the budget-balance constraint may simply fail. Thus, one of our main research questions is how we can design a market mechanism with good properties in the presence of these domain-specific challenges.

Based on the query, the buyer's value function, the sellers' cost functions, and the statistics, the market computes the estimated value for the different result sets. The buyer does not specify a value for the different result sets itself as the estimation of this value is a computationally challenging task.

1.3 Overview of Contributions

In this paper, we make the following contributions:

1. We propose the "Query Market," a double auction mechanism for querying the WoD.
2. We show that the *Threshold* rule does not work in our domain due to the uncertainty about the sellers' data.

3. We introduce three *payment correction* rules to design payment rules that always satisfy the *participation* constraint despite the uncertainty in our domain.
4. We evaluate the efficiency of the three payment rules via a computational Bayes-Nash equilibrium analysis (Lubin, Bünz and Seuken (2015)) for multiple market scenarios.

1.4 Related Work

1.4.1 Market-based Approaches towards Resource Allocation in Computational Systems

The idea to use markets to allocate computational resources is almost as old as computers themselves. Already in the 1960s, researchers at Harvard University used an auction-like method to determine who gets access to the PDP-1, the world's first interactive, commercial computer (Sutherland (1968)). Since then, many market-based approaches for computational systems have been proposed.

Early research on market-based scheduling focused on the efficiency of computational resource allocation. Malone, Fikes and Howard (1983) present Enterprise, and show how to achieve an efficient allocation of tasks between multiple LAN-connected nodes, where task processors broadcast requests for bids and bid on tasks. Bids reflecting task completion times. Likewise, Spawn by Waldspurger et al. (1992) utilizes a market mechanism to optimize the use of idle resources in a network of workstations. Van Alstyne, Brynjolfsson and Madnick (1995) investigated the impact of soft factors such as ownership for incentive-provisioning in database systems.

More recently, Lai et al. (2005) proposed Tycoon, a distributed computation cluster, featuring a proportional-share market resource allocation model. The authors claim that an economic mechanism is vital for large scale resource allocation a common problem on the Web. Furthermore, market-based optimizations have proved to be as good or better than traditional allocation methods in grid-computing schedulers. Similarly, Auyoung et al. (2006) demonstrate how profit-aware algorithms outperform non-profit aware schedulers across a broad range of scenarios.

Labrinidis, Qu and Xu (2007) applied market-based optimizations to real-time query answering systems. Stonebraker et al. (1996) proposed a WAN-scale Relational Database Management System with a market-based optimizer instead of a traditional cost-based one. Dash, Kantere and Ailamaki (2009) proposed a market-based approach for cloud cache optimization taking into account a user's value for getting an answer to a query. However, their approach focuses on the cost-side of cloud computing. Koutris et al. (2013) proposed a Market for SQL queries which sells data instead of computational resources for answering queries. They use an arbitrage-free pricing scheme instead of a double auction to calculate payments and do not consider competition between sellers in their analysis. Furthermore, buyers in their market do not specify a value for different results nor do sellers specify estimated statistics about their data, hence, they do not face the problem of value and cost estimation, one of the main issues in our market.

Only recently, a new research field called electronic market design has emerged (Anandalingam, Day and Raghavan (2005)). This field provides computer scientists with the necessary tools from auction theory, mechanism design, and market design, to analyze and design markets for computational resources with the same precision as economists have done with great success, for example, in the multi-billion dollar spectrum auction domain (Cramton (2013)). Our goal in this project is to develop a market-based approach for the WoD that is equally grounded in the mathematical foundations of mechanism design and market design.

1.4.2 Querying the Web of Data

Research on distributed query processing has a long history in the database field. Its traditional concepts were used to provide integrated access to RDF sources distributed on the WoD (Harth et al. (2007), Quilitz and Leser (2008), Erling and Mikhailov (2009)). The drawback of these

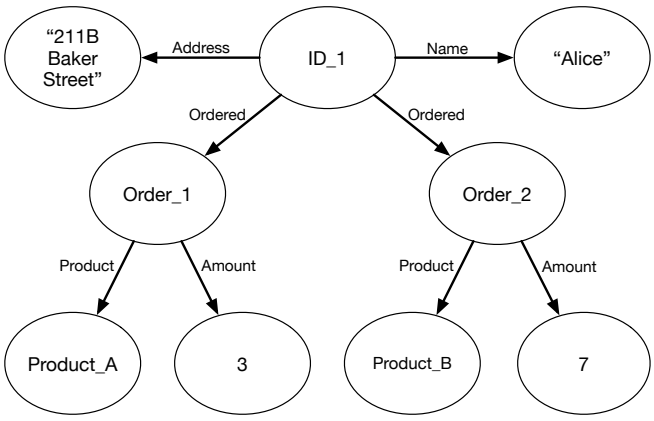

Figure 1: An example dataset, represented as an RDF graph, about a customer "Alice" who ordered 3 items of product A and 7 items of product B.

solutions is that they assume total control over the data distributions, an unrealistic assumption in the open Web. Addressing this drawback, some proposed systems that do not assume fine-grained control but perfect knowledge over the `rdf:type` predicate distribution (Langegger, Wößand Blöchl (2008)) while others proposed to extend SPARQL with explicit instructions controlling where to execute sub-queries (Zemánek, Schenk and Svatek (2007)). Unfortunately, this assumes an ex-ante knowledge of the data distribution on part of the query writer.

More recently, SPLENDID (Görlitz and Staab (2011)) relies on service descriptions and VoID statistics for each endpoint, to perform source selection and query optimisation. In contrast, FedX (Schwarte et al. (2011)) uses no knowledge about mappings or statistics about concepts/predicates but consults all endpoints to determine if a predicate can be answered (caching this information for the future). Finally, Hartig, Bizer and Freytag (2009) describe an approach for executing SPARQL queries over Linked Open Data (LoD) based on graph search. LoD rules, however, require them to place the data on the URI-referenced servers a limiting assumption for example when caching/copying data. Other flexible techniques have been proposed, such as the evolutionary query answering system *eRDF* (Oren, Gueret and Schlobach (2008)), where genetic algorithms are used to learn how to best execute a SPARQL query. However, none of these approaches investigates the economic viability of their proposed solutions.

2 Web of Data (WoD)

One of the main concepts of the WoD is the Resource Description Framework (RDF) (Cyganiak, Wood and Lanthaler (2014)) which models data as *statements* about *resources* (*subject* and *object*) which are linked via a *predicate* that defines the relation between the two resources. A resource denotes something in the universe of discourse (e.g. a physical thing, a website, or an image). Figure 1 shows how data in RDF can be modelled as a graph. Each arrow in the figure represents a predicate pointing from a subject to an object. For example the resource *ID_1* in Figure 1, representing the ID of a customer, is the subject of both, the *Address* and *Name* predicates. The objects of these relations are the literals "221B Baker Street" and "Alice", respectively.

The SQL-like query language SPARQL (Harris and Seaborne (2013)) was proposed as a language to query the Web of Data. Listing 1 shows an example of how such a query can look like. The query asks for name and address of customers, the products ordered by them and the amount. If this query is executed against the dataset in Figure 1 the result from Table 1 will be returned. Each line ending with a "." in the WHERE-clause indicates a triple pattern. The

leading "?" in the query indicates variables. During query processing the query engine searches for binding for these variables such that the result matches the corresponding RDF graph. Each valid match for the variables will produce one row in the result. In a given row, all variables with the same name must be matched to the same resource. For example if the variable "?order" is matched with the resource *Order_1* in the second line of the query in Listing 1 then the variable "?order" in the third line must also be matched with the resource *Order_1*. Hence, in this case the only valid binding for the variable "?amount" is 3 which gives us the first row in Table 1.

```
SELECT ?name ?address ?product ?amount WHERE {
        ?order Product ?product .
        ?order Amount ?amount .
        ?id Ordered ?order .
        ?id Name ?name .
        ?id Address ?address . }
```

Listing 1: A query which asks for name, address, product and amount.

?name	?address	?product	?amount
"Alice"	"221B Baker Street"	Product_A	3
"Alice"	"221B Baker Street"	Product_B	7

Table 1: Result of the query in listing 1.

2.1 Distributed SPARQL processing

If the required data for a query is not located in a single place but distributed over different datasets queries must be processed in a distributed fashion. One of the problems in distributed SPARQL processing is the estimation of the size of a join of triples from different datasets. Figure 2 shows how data could be partitioned over two datasets, A and B. If the same query from Listing 1 is performed over A and B the result will be again as in Table 1. The size of the join is in this case 2 even though there are 4 orders in dataset A and 3 customer names in dataset B. The problem of join estimation is that given some statistics from both individual datasets (in this case the count of triples) it is hard to guess the size of the join of these two datasets. It could be that every one of the four orders matches a customer in B which has 3 different addresses. In this case the result size would be 12, 3 times the same name with 3 different addresses for every order. But it could also be that the data from A and B don't match and hence the result size would be 0. In the field of distributed query processing there are several different approaches to estimate the size of the join without executing the query itself. Each approach has a different trade-off between accuracy and cost (consumption of computational resources).

3 The Query Market

The market mechanism we propose is a double auction that allows a buyer to submit a query and get an answer to his query using datasets provided by different sellers.[2]

[2]To simplify the language, we will use "he" for buyers and "she" for sellers.

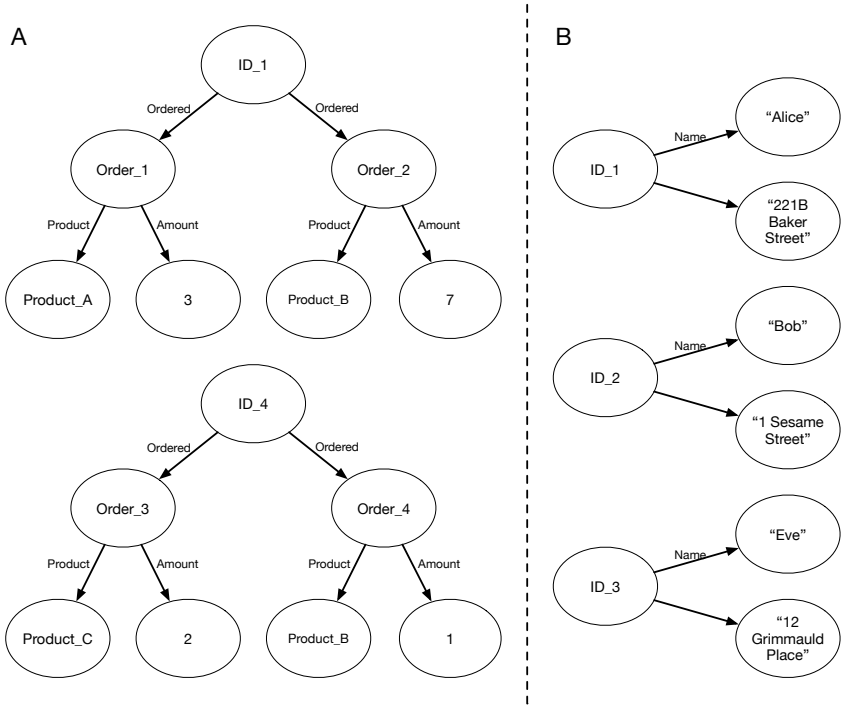

Figure 2: RDF graphs distributed over two datasets, A and B

3.1 Modeling Assumptions

We assume that sellers are only constrained by *what kind of data they can provide*, but that they are not resource-bounded, i.e. that they can potentially answer all the queries for which they have data to provide. This assumption is motivated by the fact that computational resources can nowadays be bought dynamically when needed, for example from cloud services like Amazon EC2. Hence, the cost of the sellers is the marginal cost that occurs answering an additional query which is the cost of the additional resources that need invested to answer the query.

As this implies that buyers are not competing for the same resources, we can hold an independent auction for each buyer's query. Hence, the remaining of the paper will only consider one buyer per auction. In practice multiple auctions run in parallel, one for each query.

Furthermore, we assume that there is no incentive for the sellers to strategize on the statistics they provide for their data. This assumption can be motivated in two ways: (1) the market operator may have the right to audit the statistics of sellers at any point in time, such that incorrect statistics can be detected and penalized; (2) the market operator may directly run a process on the sellers' machines to fetch the statistics in regular time intervals. However, even though the sellers are non-strategic about reporting the statistics, they are only rough (and often incorrect) estimates of what the sellers can deliver, because "precise" estimates might be to expensive to produce. Hence, the statistics delivered by the sellers might be flawed. Additionally, the result of a join between data from different sellers is again an estimate based on their statistics, and thus, imprecise estimates may compound when computing join estimates.

Finally, we assume that the buyer is only strategic about reporting his value function, but submits his true query to the market. The goal of the buyer is to maximize his utility given his query. This means maximizing his value minus the cost he needs to pay for the answer.

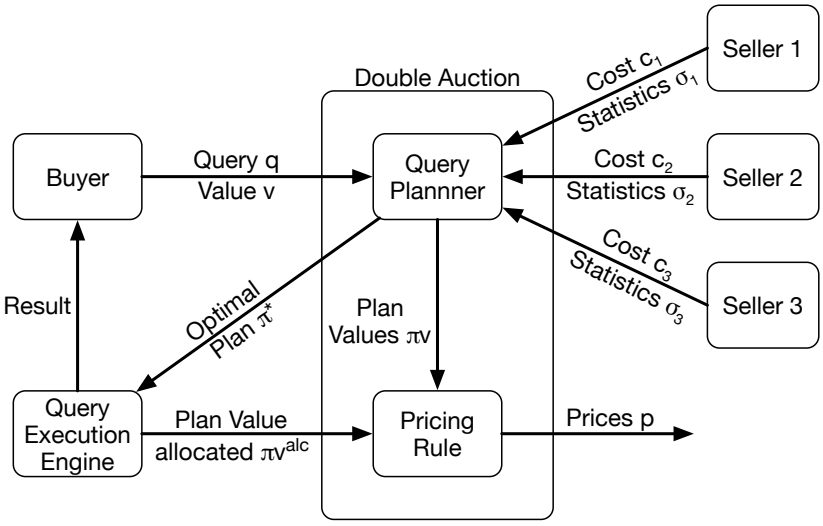

Figure 3: Schematic view of the *Query Market*.

3.2 Formal Model

Figure 3 shows a schematic of our market with three different sellers. We let b denote the buyer and $S = \{s_1, \ldots, s_n\}$ denote the set of sellers in the market. The buyer must submit the query q and the value function v. The pair (q, v) of query and value function is the bid of the buyer. Each seller i needs to provide some statistics σ_i about the data she can provide for the query as well as her cost function c_i. The pair (σ_i, c_i) of statistics and cost function is the bid of seller i.

Example 1. *Consider the query from Listing 1. An example of a value function v would be the number of rows in the result times some constant value, e.g. \$5. In this case the value function would look like:*

$$v = valuePerRow(r) := r \times \$5$$

where r is the number of rows in the result.

For an allocation that produces the result from Table 1 the value would be \$10.

Analogously, a seller can for example specify her cost with respect to the number of triples provided, e.g. \$2. In this case the cost-function c_i would look like:

$$v = costPerTriple(t_i) := t_i \times \$2,$$

n where t_i is the number of triples provided by seller i.

If seller i would provide for a certain allocation 10 triples, her cost would be \$20.

A plan π describes how the query should be answered by the *Query Execution Engine*. This includes the set of sellers that participate in answering the query as well as the concrete instructions for the *Query Execution Engine* how to produce the result for the query.

The *Query Planer* constructs all possible, valid plans that can answer the query q. For each plan π the *Query Planer* calculates the *return set summary* r_i^{est} which is an estimate of what seller i will return if plan π is executed (e.g. the cardinality of the data the seller returns) and combines all the required information to calculate the cost for seller i. Formally, r_i^{est} is the result of r_i which denotes the estimation process and is a function of the plan π and the statistics $\sigma_1, \ldots, \sigma_n$ provided by the sellers:

$$r_i^{est} := r_i(\pi, \sigma_1, \ldots, \sigma_n).$$

The *Query Planer* also calculates the *result set summary* x^{est} which is an estimate of what the buyer will get if plan π is executed (e.g. the cardinality of result of the query) and combines all the required information to calculate the value for the buyer. Note that x^{est} does not denote the actual result of the query as it is not needed for the pricing rule. Formally, x^{est} is the result of x which denotes the estimation process and is a function of the plan π and the statistics $\sigma_1, \ldots, \sigma_n$ provided by the sellers:

$$x^{est} := x(\pi, \sigma_1, \ldots, \sigma_n).$$

Using x^{est} and r_i^{est}, the estimated value $v^{est} := v(x^{est})$ for the buyer and the estimated cost $c_i^{est} := c_i(r_i^{est})$ for seller i can be calculated. These estimates, together with the actual plan π form the vector of *plan values* πv:

$$\pi v := (\pi, v^{est}, c_1^{est}, \ldots, c_n^{est}).$$

Example 2. *Consider a plan π that consists of a join between two sellers, s_1 and s_2. Based on the statistics σ_1 and σ_2, the market estimates x^{est}, r_1^{est}, and r_2^{est}. Given the value and cost functions from example 1, assume that for the join in π seller s_1 needs to provide 10 triples, seller s_2 5 triples, and the cardinality of the result is 8 rows. In this case we have:*

$$r_1^{est} := r_1(\pi, \sigma_1, \sigma_2) = 10,$$

$$r_2^{est} := r_2(\pi, \sigma_1, \sigma_2) = 5,$$

$$x^{est} := x(\pi, \sigma_1, \sigma_2) = 8.$$

With these estimates, the vector of plan values πv can be formed:

$$\pi v := (\pi, v^{est} = \$40, c_1^{est} = \$20, c_2^{est} = \$10).$$

The set of all plan values will be sent to the *Pricing Rule*, and it contains all the information that is required by the *Pricing Rule* to calculate the prices.

At the same time, the *Query Planer* determines the plan π^* that maximizes the estimated social welfare

$$\pi^* = \underset{\pi}{\operatorname{argmax}}\,(SW(\pi, \sigma_1, \ldots, \sigma_n)) = \underset{\pi}{\operatorname{argmax}}\Big(\underbrace{v(\underbrace{x(\pi, \sigma_1, \ldots, \sigma_n)}_{x^{est}})}_{v^{est}} - \sum_i \underbrace{c_i(\underbrace{r_i(\pi, \sigma_1, \ldots, \sigma_n)}_{r_i^{est}})}_{c_i^{est}}\Big)$$

and sends the allocated plan to the *Query Execution Engine*[3].

The execution of the query produces the result which is sent to the buyer. Using the result of the query execution both, the estimate of the *result set summary* x^{est} and the estimate of the *return set summaries* r_i^{est} of the allocated plan, can be replaced by the actual summaries x^{alc} and r_i^{alc}, where "alc" stands for "allocated". The actual value $v^{alc} := v(x^{alc})$ and costs $c_i^{alc} := c_i(r_i^{alc})$ for the result can now be computed and the vector of the allocated *plan values* $\pi v^{alc} := (\pi^*, v^{alc}, c_1^{alc}, \ldots, c_n^{alc})$ is sent to the *Pricing Rule*.

Example 3. *Assume that the plan from example 2 is the allocated plan. During query execution it might turn out that seller s_1 needs to provide 10 triples for the join but seller s_2 only 4. At the same time, the result set consists only of 5 rows instead of 8. Hence, the vector of the allocated plan values is:*

$$\pi v^{alc} := (\pi^*, v^{alc} = \$25, c_1^{alc} = \$20, c_2^{alc} = \$8).$$

The *Pricing Rule* (further described in Section 4.2) computes the price for the buyer, p_b, and the prices for the sellers, $p_1, \ldots, p_n$. We assume that both, the buyer and the sellers, have a quasi-linear utility function. For the buyer, the utility function is given as follows:

$$u_b(x) = v(x) - p_b \qquad \text{for } x \in \{x^{est}, x^{alc}\}$$

[3]The execution is done in a distributed fashion against the set of sellers from the winning plan using the Federated Query Extension of SPARQL 1.1. See http://www.w3.org/TR/sparql11-federated-query/

and for the sellers as follows:

$$u_i(r) = c(r) - p_i \qquad \text{for } r \in \{r^{est}, r^{alc}\}$$

Remark 1. *Even though the buyer himself does not explicitly specify different bids for different plans, the auction implicitly translates his value function v into a specific bid $v(\pi, \sigma_1, \ldots, \sigma_n)$ which is the value for a given plan π and statistics $\sigma_1, \ldots, \sigma_n$. Since each plan generally consists of multiple sellers, the buyer's value function eventually translates into combinatorial bids for different bundles of sellers. Thus, our auction is a combinatorial double auction.*

4 The Double Auction

When we are designing a double auction, we strive for the following four properties:

1. Efficiency (EFF)
2. Budget-balance (BB)
3. Participation (P)
4. Strategyproofness (SP)

It is well-known, that in an exchange domain, we cannot satisfy all four properties simultaneously. In our domain, we relax strategyproofness and aim for a double-auction that maximizes efficiency, subject to satisfying budget-balance and participation.

4.1 Allocation Rule

The allocation is part of the Query Planner (as described in Section 3) which chooses a plan π^* that maximizes the estimated social welfare $SW = v^{est} - \sum_i c_i^{est}$.

4.2 Pricing Rule

For this section, we generalize and unify the notation for buyers and sellers and denote the set of all *agents* as $I = S \cup \{b\}$. We use index 0 for the buyer and $1...|S|$ for sellers. We let a^* denote an efficient allocation (which corresponds to the plan π^*) and a^*_{-i} an efficient allocation when agent i is excluded. Now, let $w_j(a)$ denote an agent's value for an allocation a, $j \in 0, 1, ..., |S|$, where $w_0(a) = v(a)$ for the buyer, and $w_j(a) = -c_j(a)$ for a seller $j = \{1, ..., |S|\}$. The social welfare for the efficient allocation including all agents is denoted $V^* := \sum_j w_j(a^*)$, and the social welfare for the efficient allocation excluding agent i is denoted $V^*_{-i} := \sum_{j \neq i} w_j(a^*_{-i})$. VCG payments can be computed as follows:

$$p_i^{vcg} := w_i(a^*) - (V^* - V^*_{-i}).$$

We let $\Delta_i^{vcg} := V^* - V^*_{-i}$ denote the *VCG discount* of agent i. Furthermore, we let $\Delta^{vcg} := \{\Delta_0^{vcg}, ..., \Delta_{|S|}^{vcg}\}$. Now, we can equivalently write VCG payments as follows:

$$p_i^{vcg} := w_i(a^*) - \Delta_i^{vcg}.$$

At first sight, VCG payments may seem attractive, because they make truthful reporting of values, a dominant strategy for all agents. However, it is well known that in two-sided market setting like ours, VCG generally leads to a budget-deficit, i.e., the payments made to the sellers can be larger than the payments collected from the buyers (Milgrom, 2007). Formally, the budget-balance constraint is violated.

To address this problem, Parkes, Kalagnanam and Eso (2001) have proposed a family of pricing rules that only *approximate* VCG payments, while ensuring budget-balance. Their idea is to compute payments as $p_i := w_i(a) - \Delta_i$, where Δ_i is a discount assigned to an agent for allocation

a and $w_i(a)$ is a reported value of the agent. Discounts $\Delta = \{\Delta_0, ..., \Delta_{|S|}\}$ are computed to minimize the distance to VCG discounts while satisfying the budget-balance and participation constraints. The *Threshold* pricing rule (Parkes, Kalagnanam and Eso, 2001) can be formulated as the following optimization problem, where the target point is $\Delta^{trg} = \Delta^{vcg}$:

$$\begin{aligned} &\min_{\Delta} L_2(\Delta, \Delta^{trg}) \\ &s.t. \sum_{i \in \{b\} \cup S} \Delta_i = V^* \qquad (budget - balance) \\ &0 \leqslant \Delta_i \leqslant \Delta_i^{trg} \ \ \forall i \in I \qquad (participation) \end{aligned} \tag{1}$$

4.3 Failure of the Participation Constraint

The *Threshold* pricing rule provides budget-balance in a two-sided market domain with no uncertainty about the goods to be sold. However, in our domain, because of inefficiencies which can occur due to uncertainty in statistics provided by the sellers about their data, all of the pricing rules proposed by Parkes, Kalagnanam and Eso (2001) may lead to an infeasible participation constraint in the payment computation problem.

We now provide an example which illustrates how imprecise estimates provided by the sellers can cause problems with the participation constraint. In particular, the following example shows that even VCG can end up violating the participation constraint in a domain with uncertainty.

Example 4. *Let b be the buyer and $S = \{s_1, s_2, s_3\}$ three different sellers. Assume two possible plan values are generated by the Query Planer for b's query. The first plan value $\pi v_1 = \{\pi_1, v^{est}(\pi_1) = 3, c_1^{est}(\pi_1) = 0.5, c_2^{est}(\pi_1) = 2\}$, the second one $\pi v_2 = \{\pi_2, v^{est}(\pi_2) = 6, c_2^{est}(\pi_2) = 1, c_3^{est}(\pi_2) = 4\}$. If there was no uncertainty in the domain, an efficient allocation would choose πv_2 resulting in VCG payments $p_b^{vcg} = 5$, $p_{s_1}^{vcg} = 0$, $p_{s_2}^{vcg} = -2$, $p_{s_3}^{vcg} = -4.5$. While not being budget-balanced, these payments provide participation.*

Now assume that there is an uncertainty and after πv_2 is allocated we have: $\pi v_1 = \{\pi_1, v^{est}(\pi_1) = 3, c_1^{est}(\pi_1) = 0.5, c_2^{est}(\pi_1) = 2\}$, $\pi v_2^{alc} = \{\pi_2, v^{alc} = 3, c_2^{alc} = 1, c_3^{alc} = 2\}$. VCG payments computed for this actual allocation are $p_b^{vcg} = 3$, $p_{s_1}^{vcg} = 0$, $p_{s_2}^{vcg} = -1$, $p_{s_3}^{vcg} = -1.5$. It's clear now that the participation constraint for s_3 is not satisfied anymore because $|p_{s_3}^{vcg}| < c_3^{alc}$.

If VCG violates the participation constraint, then payment rule (1) becomes infeasible in our domain. Thus, we need to design payment correction rules to tackle this problem.

4.4 Payment Correction Rules

We now propose three different payment correction rules and study which one leads to higher efficiency (in equilibrium). As mentioned before, because of uncertainty, our mechanism might sometimes make an inefficient allocation (even given truthful value reports). In this case, one cannot rely on VCG because of negative VCG discounts which lead to infeasibilities in the participation constraints (as explained in Section 4.3). A similar problem was faced by Goetzendorf et al. (2015), where the authors just trimmed such infeasibilities. We could try to apply this idea to our domain and use the trimmed VCG discounts as a target point Δ^{trg} in (1). While this new target guarantees satisfaction of the participation constraint, it is not clear whether it provides good incentives (and thus good efficiency). For this reason, we also define two alternative methods for constructing a target point Δ^{trg}. For this we let π^* denote the plan for the allocation chosen by the Query Planer. We the define $V_{est}^* := v(x(\pi^*, \sigma_1, \ldots, \sigma_n)) - \sum_i c_i(r(\pi^*, \sigma_1, \ldots, \sigma_n))$ as the total value of all agents in this allocation (before query execution), and $V_{est,-i}^* := v(x(\pi_{-i}^*, \sigma_1, \ldots, \sigma_n)) - \sum_i c_i(r(\pi_{-i}^*, \sigma_1, \ldots, \sigma_n))$ as the total value of all agents in the allocation where agent i is excluded. Our three payment correction rules are:

1. `PC-TRIM`: Compute VCG discounts using x^{alc} and r^{alc} for the allocated plan, given a possibly non-efficient allocation (similarly to Goetzendorf et al. (2015)). Then trim infeasibilities in the participation constraints by increasing all negative Δ_i to be zero and use the resulting discounts as a new target point Δ^{trg}.
2. `PC-PENALTY`: We let $V^*_{alc,i}$ denote the social welfare for the allocated plan π^*, where we use actual summaries (based on the result from the executed query) for seller i, but use the reported statistics for all other sellers. Then we compute VCG discounts based only on the statistic reported by sellers. Next we reduce those discounts which correspond to sellers who have provided wrong estimates by a penalty factor $\alpha_i = \frac{V^*_{est} - V^*_{alc,i}}{V^*_{est}}$, for all i involved in π^*. The penalty factor reflects the harm caused by agent i to the efficiency by providing her imprecise estimates.
3. `PC-VCG`: Compute VCG discounts using the (possibly wrong) estimates x^{est} and r^{est} from the Query Planner, instead of x^{alc} and r^{alc}, even though the query has already been executed, and use those estimate-based VCG discounts as a target point Δ^{trg}.

The resulting payment rules for our market are derived from the *Threshold* payment rule by applying the aforementioned payment correction procedures. However, it is not strictly necessary to use *Threshold* as an underlying payment computation rule. Instead many other VCG approximation techniques can be utilized in combination with the proposed payment correction rules, for example, *Small*, *Large*, *Fractional* (Parkes, Kalagnanam and Eso (2001)) etc.

To check which payment rule provides better efficiency, we compare all three of them in a double auction setup when all agents play according to the Bayes-Nash equilibrium (BNE) of the mechanism.

5 Quantitative Evaluation

To study the efficiency of the three proposed payment rules, we need to make an assumption regarding the agents' behavior in the market. Assuming agents to be rational, we allow them to strategize and act as if they would try to maximize their expected utilities. This allows us to naturally restrict our analysis to only setups when all agents play according to their *equilibrium strategies*. In other words, we are interested in finding a strategy profile for all agents which would maximize the expected utility of every agent assuming that all other agents also play according to the strategies in this strategy profile. This approach allows us to capture the strategic behavior of the agents (due to non-strategyproof payment rules) and evaluate efficiency loss caused by the agents' strategic behavior. Given that a full information assumption seems highly unrealistic in this domain, we adopt a *Bayes-Nash equilibrium* analysis approach.

However, theoretical evaluation of the BNE can be an extremely difficult task. Goeree and Lien (2014) derived a BNE for a simple combinatorial auction setup with only three agents and two items. Ausubel and Baranov (2010) derived a BNE for a similar setup but taking into account correlated values of agents. However, there are no any further results for neither more complicated combinatorial auctions nor for double auctions. In addition to that, our market usually involves much more than just 3 agents which makes the theoretical analysis even harder. Thus, instead of deriving the equilibrium analytically we try to approximate it computationally. Section 5.1 describes more precisely how a BNE can be approximated. Section 5.2 introduces the simulated WoD domain and describes the quantitative evaluation framework in detail. In conclusion Section 5.3 provides a detailed explanation of the results of our evaluations.

5.1 Approximate BNE Computation

To perform the quantitative evaluation of our market we compute (ϵ, δ)-BNEs using a method described by Lubin, Bünz and Seuken (2015). This method is based on an iterative best-response search procedure.

The general approach for BNE approximation implements an idea of a *Fictitious Play* and can be described as a two stage procedure. First, agents are grouped into one of n_b bins. We use only $n_b = 2$ bins: one for buyers and one for sellers. Agents within same bin are assumed to play the same strategy. We also assume that the strategy space for a buyer with value v can be represented as $v \cdot s, s \in [0, 1]$ implying that the buyer can only underbid but not overbid. Similarly, the strategy space for a seller with a cost c is $c \cdot \frac{1}{s}, s \in [0, 1]$ which means that the seller only overreport his costs.

The second stage is an iterative process during which for every bin a best response strategy is computed, i.e., a strategy which is when being played by every agent within the bin maximizes the total utility of all agents in the bin (assuming the strategies of other agents are fixed). At this stage on every iteration `nSamples` games are sampled using known distributions of types of all agents. Then, each agent within a particular bin plays different strategies while strategies of all other agents are fixed. Every such a game contributes into the overall utility of the bin which is computed as a sum of utilities of all agents from the bin. This reduces the aforementioned problem of the best response strategy search to a global stochastic optimization problem where the objective is to maximize the total utility for every bin while varying strategies of agents within the bin. Clearly, the number of such global optimization problems per iteration is equal to the number of bins. When optimal strategies for bins are computed, we apply them for every agent in corresponding bins by allowing them to play a convex combination of their current strategy and the newly found best response. This leads to an update of the strategy, or *shaving factor*, for every bin and followed by the next iteration.

The process terminates when an (ϵ, δ)-BNE is identified, i.e., when playing the best response rather than the (ϵ, δ)-BNE strategy does not improve bins' overall utilities more than by a factor of $1 + \epsilon$. More formally,

Definition 1. *A strategy profile $s^* = (s_1^*, ..., s_{nBins}^*)$ is an (ϵ, δ)-Bayes-Nash equilibrium if for every bin $i \in \{1, ..., nBin\}$*

$$\frac{\mathbb{E}_{-i}[u_i(br_i, s_{-i}^*)]}{\mathbb{E}_{-i}[u_i(s_i^*, s_{-i}^*)]} \leqslant 1 + \epsilon,$$

and

$$||s_i^* - br_i(s_{-i}^*)||_2 \leqslant \delta,$$

where br_i denotes a best response of the i-th bin and u_i is a total utility of the i-th bin.

We use $\epsilon = 1\%$, $\epsilon = 2.5\%$ and $\epsilon = 3\%$ for the scenario with three, five and ten sellers respectively. For all scenarios we assume $\delta = 0.1$. We also varied `nSamples` from 1000000 for a small setup with only three and five sellers to 500000 for the setup with 10 sellers. To solve the described global optimization problem of a best response search we utilize multistart method with 10 starting points generated randomly from $U[0, 1]$ combined with a pattern search local optimization procedure.

5.2 Domain Description

To evaluate the market we generate three different WoD domains and thus simulate three different market scenarios. The first scenario includes a single buyer b_1 and three sellers s_A, s_B and s_C. For this scenario we assume that the Query Planner always generates two plans π_1 and π_2. Each plan involves two different sellers chosen randomly from a uniform distribution. We denote $t_X^{est}(\pi_i)$ the estimated (and possibly imprecise) number of triples which a seller s_X reports she can provide for the plan π_i, $X \in \{A, B, C\}$, $i \in \{1, 2\}$. The numbers of triples provided by different sellers is chosen uniformly from $[1, 10]$. Even though such a case with only three sellers is not very likely to happen in a real Query Market where the number of sellers is usually large, it brings out some interesting properties of the market. More specifically, we use this scenario to study what happens with the market if it has very influential agents.

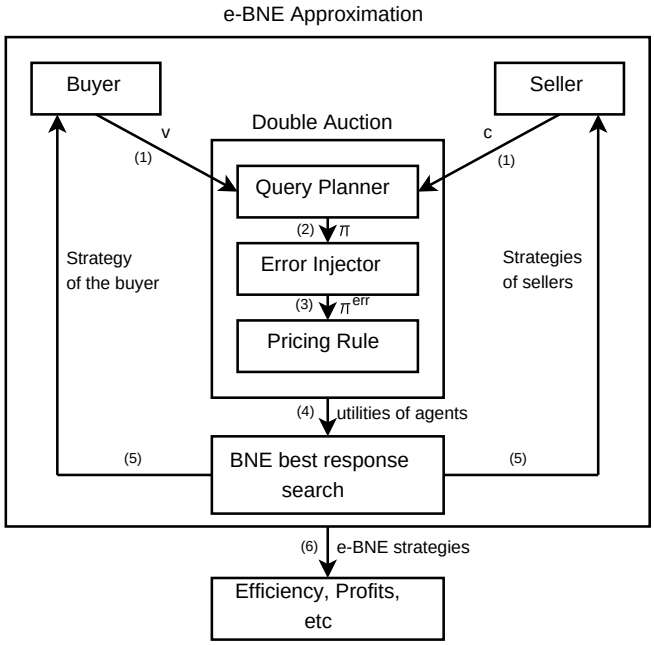

Figure 4: Evaluation framework.

The second scenario includes a single buyer and five sellers. For this scenario we again generate two different plans each involving two different sellers chosen randomly from a uniform distribution. By comparing the resulting strategies of this scenario with the corresponding strategies from the previous one, we study how the factor of the number of sellers alone influences on the efficiency of our market.

The third scenario is more typical for the Query Market and includes a single buyer and 10 sellers. For this scenario we generate five different plans. Similarly to previous two scenarios each plan involves two different sellers chosen randomly from a uniform distribution. A high diversity of sellers and a small number of sellers involved in a single plan reduces the effect of individual sellers.

For all domains, we assume the agents' value and cost functions are linear in the cardinality of data. Thus, we only need a value and cost factors, v and c_i, $i \in S$ respectively to be reported by agents. We assume that v and c_i are drawn independently from a uniform distribution $U[0,1]$, $(i \in S)$.

When a plan is allocated an estimation error can be injected into the plan. To model estimation errors (or uncertainty), we randomly select a seller $X \in \{A, B, C, ...\}$ from a uniform distribution over the set of sellers involved in the allocated plan π^*. We then generate $t_X^{alc} \sim \mathcal{N}(t_X^{est}, \sigma)$ truncated at 0 instead of t_X^{est}. Furthermore, we assume that a seller can only overestimate t_X^{est}, i.e., $t_X^{est} \geqslant t_X^{alc}$; and if $t_X^{alc} > t_X^{est}$, we use $t_X^{alc'} = t_X^{est} - (t_X^{alc} - t_X^{est})$ instead. Underestimation of the number of allocated triples should not be a problem as in this case the seller can simply discard extra triples which are not required by the plan. Additionally, by varying σ we study how the degree of uncertainty in the domain influences on agents' incentives and the resulting market efficiency.

For the three described scenarios, we first approximate the Bayes-Nash equilibrium for every payment rule. Then we feed the mechanism with all agents playing according to the BNE into a benchmarking module to measure the following characteristics:

Type of statistic	Payment rule	1 Buyer, 10 Sellers							
		Strategies, %			% of VCG profits		rate of BB	BB	Partici-
		Buyer	Sellers	Efficiency	Buyer	Sellers	violations,%	deficit	pation
w/o uncertainty	VCG	100	100	1.00	100	100	72	0.87	✓
	THRESHOLD-TRUTH	100	100	1.00	69	76	0	0	✓
	THRESHOLD	76	147	0.39	48	42	0	0	✓
w. uncertainty	VCG-TRUTH	100	100	0.86	89	80	66	0.80	×
	VCG	98	109	0.77	80	77	64	0.89	×
	PC-TRIM	78	142	0.37	43	37	1.8	0.005	✓
	PC-PENALTY	84	135	0.46	46	44	1.6	0.003	✓
	PC-VCG	80	136	0.41	45	41	1.7	0.004	✓

Table 2: Results of a computational BNE analysis for all pricing rules in a uncertainty-free and uncertain setting for the scenario with one Buyer and 10 Sellers.

- The total efficiency of the market relative to the total efficiency of the market if VCG is used to compute payments and there is no uncertainty in the domain.
- Profits of buyers and profits of sellers relative to their profits if VCG is used for payment computation in a domain without uncertainty.
- The rate of budget-balance violations, i.e., how often does it happen that the budget-balance constraint is violated.
- The amount of budget-balance deficit w.r.t. the total efficiency of the market, i.e., the ratio of the budget-balance deficit to the total social welfare.

The general workflow of our evaluation framework is presented in Figure 4. First, buyer's and sellers' types are sampled and their value and cost functions are submitted into the auction (1). Second, a set of plans is generated by the Query Planner (2). Then an allocation happens followed by an injection of an estimation error into the allocated plan (3). Payments are computed on the next step. Resulting utilities of agents are used in a best response search procedure to identify optimal strategies for agents on the current iteration. The process repeats (5) using strategies computed on the previous iteration until ϵ-BNE is found and then these BNE strategies are fed into the efficiency, profits and budget-balance evaluation module (6). The efficiency module in its turn generates 100,000 markets with agents acting according to the identified BNE and computes the total efficiency, profits of agents, number of budget-balance constraint violations and the budget amount required to cover these violations etc.

5.3 Results

Table 2, Table 3 and Table 4 show our evaluation results. We have considered two different cases (with and without uncertainty): within the uncertainty-free setting, we consider `VCG`, `THRESHOLD-TRUTH`, and the standard `THRESHOLD` rule. The `THRESHOLD-TRUTH` rule is simply the standard Threshold rule where instead of playing their BNE strategies, we force all agents to play truthfully. While being very artificial, this mechanism gives us an upper bound on the agents' profits for a budget-balanced mechanism in the uncertainty-free setting.

For the domain with uncertainty, we consider five pricing rules: `VCG-TRUTH` (a benchmark which obtains true values, but uncertain reports from sellers and uses VCG as a payment rule), `VCG` (a benchmark which computes VCG using uncertain reports), and then the three price correction (PC) rules `PC-TRIM`, `PC-VCG` and `PC-PENALTY`. Note that `VCG` for the setup with uncertainty is not truthful anymore due to inefficient allocations and thus the agents will play their BNE strategies.

5.3.1 Market Scenario #1: 1 Buyer and 10 Sellers

Consider Table 2, where we present the results for market scenario #1 with 1 buyer and 10 sellers. This scenario is meant to represent a typical WoD domain with lots of sellers, where no individual seller is pivotal for trade to happen.

The first observation we make is that `THRESHOLD` only achieve 39% of the efficiency of `VCG`. This result is not surprising, because while `VCG` is strategyproof, `THRESHOLD` is not. As we can see, under `THRESHOLD`, the agents start manipulating (the buyer under-reports by 24% and the sellers over-report by 47%), which results in lost trades, and thus explains the lower efficiency of `THRESHOLD`.

Next, we compare `VCG` in the domain without uncertainty to `VCG-TRUTH` in the uncertain setting. Here, the 14% smaller efficiency of `VCG-TRUTH` indicates how much efficiency we lose due to imprecise estimates and not because of agents' manipulations. Then we compare `VCG` with `VCG-TRUTH` in the domain with uncertainty. Here we observe that `VCG` loses an additional 9% points over and above the efficiency loss that `VCG-TRUTH` already incurred, which must now partly be due to the strategy manipulations of the agents. Added together, the uncertainty injection and the agents' strategic manipulations lead to a 23% efficiency loss, when compared to `VCG` in the uncertainty-free setting.

Finally, we consider our three payment correction rules, `PC-TRIM`, `PC-PENALTY` and `PC-VCG`, whose efficiency is between 0.37 and 0.46. Let's first compare the resulting efficiency to that of `VCG-TRUTH`, the efficiency upper-bound for the market with uncertain reports. For example, `PC-VCG` achieves 48% of the efficiency that `VCG-TRUTH` achieves. Now let's consider `PC-TRIM`, which actually uses `VCG` from the setting with uncertainty as its target point in the price computation algorithm. Thus, the efficiency of `PC-TRIM` is upper-bounded by the efficiency of `VCG` in the setting with uncertainty. Here, we obtain that `PC-TRIM` achieves 48% of the efficiency of `VCG` in the setting with uncertainty. Note that it is not surprising that in some cases (like for `PC-TRIM`) our payment correction rules do not achieve the same share of the maximum achievable efficiency that `THRESHOLD` is able to achieve in the uncertainty-free setting, because (1) we need to correct the participation constraint, which may lead to a further efficiency loss, and (2) in the domain with uncertainty, we are already starting with a target point $\Delta^{trg} = \Delta^{vcg}$ which already leads to significant strategy behavior by the agents and an efficiency loss of $(1 - \frac{0.77}{0.86}) \cdot 100\% = 11\%$. This suggests that the choice of Δ^{vcg} as a target point in the uncertain environment might be sub-optimal - in contrast to the uncertainty-free setting, where generally using Δ^{vcg} as the target point provides good incentives and thus leads to high efficiency. This effect can be seen more clearly in the scenario with smaller number of agents (see, for example, section 5.3.3).

We now look at the profits of the agents in the market, which in Table 2 are normalized relative to the profits the agents obtain under VCG in the uncertainty-free setting. We see that the buyer obtains between 43% and 46% of his VCG profits, while the sellers obtain between 37% and 44% of their VCG profits. Importantly, all three payment correction rules lead to almost identical profits. Furthermore, we see that under `PC-TRIM`, the buyer obtains a profit of $\frac{0.43}{0.89} \cdot 100\% = 48\%$ of the maximal attainable profit under `VCG-TRUTH`; and the seller obtains a profit of $\frac{0.37}{0.80} \cdot 100\% = 46\%$ of the maximum attainable profit under `VCG-TRUTH`. Thus, these results are relatively close to the 48% and 42% which `THRESHOLD` achieves in the setting without uncertainty.

5.3.2 Market Scenario #2: 1 Buyer and 5 Sellers

In Table 3 we present the results for market scenario #2 with 1 buyer and 5 sellers. This scenario represents an intermediate case between the most typical for the semantic web scenario #1 and an almost extreme scenario #3 with very influential agents.

In this case we obtained strictly worse strategies for the buyer and sellers comparing with

Type of statistic	Payment rule	1 Buyer, 5 Sellers							
		Strategies, %			% of VCG profits		rate of BB violations,%	BB deficit	Participation
		Buyer	Sellers	Efficiency	Buyer	Sellers			
w/o uncertainty	VCG	100	100	1.00	100	100	60	1.51	✓
	THRESHOLD-TRUTH	100	100	1.00	47	73	0	0	✓
	THRESHOLD	67	172	0.16	28	15	0	0	✓
w. uncertainty	VCG-TRUTH	100	100	0.86	86	85	58	1.45	×
	VCG	97	102	0.79	85	78	56	1.48	×
	PC-TRIM	72	166	0.18	26	16	1.2	0.01	✓
	PC-PENALTY	72	161	0.19	28	16	1.3	0.01	✓
	PC-VCG	71	163	0.18	27	16	1.2	0.01	✓

Table 3: Results of a computational BNE analysis for all pricing rules in a uncertainty-free and uncertain setting for the scenario with one Buyer and 5 Sellers.

those of scenario #1. This shows that the higher number of agents and available plans can considerably decrease manipulability of the market and raise its efficiency. In our case the low number of agents results in more than halved efficiency for all proposed payment correction rules. For example, we have $49\% = \frac{0.18}{0.37} \cdot 100\%$ of efficiency drop for PC-TRIM, $41\% = \frac{0.19}{0.46} \cdot 100\%$ for PC-PENALTY, and $44\% = \frac{0.18}{0.41} \cdot 100\%$ for PC-VCG. However, such a dramatic drop in efficiency is not specific only for our payment correction rules: our experiments also provide a similar trend for the THRESHOLD rule in the setup with no uncertainty. In this case we have $42\% = \frac{0.16}{0.38} \cdot 100\%$ smaller efficiency when decreasing the number of agents and available plans.

Perhaps the most surprising result is the fact that the efficiency of all payment correction rules in the setting with uncertainty (around 18%) is higher than the efficiency of THRESHOLD in the uncertainty-free setting (16%). This finding is not easy to explain, because we would typically expect that the uncertainty in the domain leads to a *reduction* in the efficiency, as we have observed it in the previous scenario. Our only explanation is that THRESHOLD, even though it is using VCG discounts as a target point, is only the optimal payment rule in an ex-post sense. However, in a Bayesian setting, only little is known about which payment rule is best (Lubin and Parkes (2009)). Our current conjecture is that our payment correction rules have coincidentally picked out such an attractive target point that, evaluated in BNE, it leads to even higher efficiency than THRESHOLD does in the uncertainty-free setting. However, further analysis is needed to evaluate this conjecture.

5.3.3 Market Scenario #3: 1 Buyer and 3 Sellers

Now consider Table 4, where we present the results for market scenario #3. Remember that this is an extreme scenario where every seller is very influential; ever seller can potentially be pivotal; and where one of the three sellers is required for all of the plans (i.e., he is a "monopolist").

The most important observation from Table 4 is that the efficiency of all three payment correction rules is now only around 15%, and thus much lower than in the scenario with 10 sellers (where it was around 40%). This can of course be explained by the fact that in this scenario, the sellers have become much more pivotal, which has increased their incentives to manipulate. While previously, the sellers where overbidding by approximately 40%, they are now overbidding by at least 63% and up to 75% (for PC-PENALTY). Remember that PC-PENALTY only penalizes sellers, and in particular those sellers that have provided imprecise estimates that have led to an inefficient allocation. Thus, the buyers can usually expect to receive high VCG discounts under this rule, and consequently play almost truthfully. The sellers, in contrast, have very large incentives to manipulate.

Type of statistic	Payment rule	1 Buyer, 3 Sellers							
		Strategies,%			% of VCG profits		rate of BB	BB	Partici-
		Buyer	Sellers	Efficiency	Buyer	Sellers	violations,%	deficit	pation
w/o uncertainty	VCG	100	100	1.00	100	100	59	1.55	✓
	THRESHOLD-TRUTH	100	100	1.00	45	73	0	0	✓
	THRESHOLD	71	178	0.17	26	14	0	0	✓
w. uncertainty	VCG-TRUTH	100	100	0.96	95	96	59	1.52	×
	VCG	98	112	0.80	84	82	59	1.54	×
	PC-TRIM	67	163	0.15	25	12	1.1	0.013	✓
	PC-PENALTY	70	175	0.15	22	11	1.1	0.014	✓
	PC-VCG	69	163	0.16	25	13	1.2	0.012	✓

Table 4: Results of a computational BNE analysis for all pricing rules in a uncertainty-free and uncertain setting for the scenario with one buyer and three sellers.

5.3.4 Budget-balance Violations and Budget Deficits

Remember that our goal was to design a budget-balanced double auction mechanism. As mentioned before, it is well known that VCG is not budget-balanced in a two-sided market. We can now put a number on how severe the budget-balance violations of VCG are in our simulated domains. In Table 2, we see that in the uncertainty-free setting, VCG led to a budget-balance violation in about 72% of the cases. In Table 4, VCG led to a budget-balance violation in about 59% of the cases.

The THRESHOLD rule is explicitly designed to be budget-balanced, but only works in the uncertainty-free setting. For this reason, we need our new price correction rules, i.e., the only rules which are always guaranteed to satisfy the participation constraint in the uncertain setting. However, even those three rules can sometimes still lead to a budget-balance violation when a plan is allocated whose total costs are larger than its value, which may happen due to imprecise estimates. These kinds of budget-balance violations cannot be avoided in our market due to the nature of the WoD. But fortunately, as we can see in Table 2 and 4, such budget-balance violations occur very infrequently for our payment correction rules: between 1.6% and 1.8% in the market with 10 sellers, and between 1.1% and 1.2% in the market with 3 sellers.

Finally, when these budget-balance violations do occur, then the amount of budget deficit that our payment correction rules exhibit is minimal. For example, in Table 2, we see that the average budget deficit for our rules is between 0.003 and 0.005, which is almost negligible compared to a budget deficit of 0.87 of VCG (all of these numbers are relative to total social welfare). In Table 4 the average budget deficit of our rules is between 0.012 and 0.014, compared to a average budget deficit of 1.55 of VCG.

Of course, in practice, we would have to find some way to cover these budget deficits. Given the almost negligible size of the budget deficit, a practical approach is to simply increase the payments charged to buyers by a correspondingly small amount on all other trades where no budget-balance violations occur. Developing a practical method for automatically calibrating these additional "taxes" such that the market operator always has enough budget to cover the expected budget deficit is subject to future work.

5.3.5 Impact of the Degree of Uncertainty on the Market Efficiency

By analyzing Table 2 and Table 3 we see that the efficiency of the proposed payment rules in the domain with uncertainty is usually higher than the efficiency of a *Threshold* rule in a domain without uncertainty. At the same time agents in an uncertain domain manipulate less than in the case when all estimates are precise. This can be caused either by a more attractive target

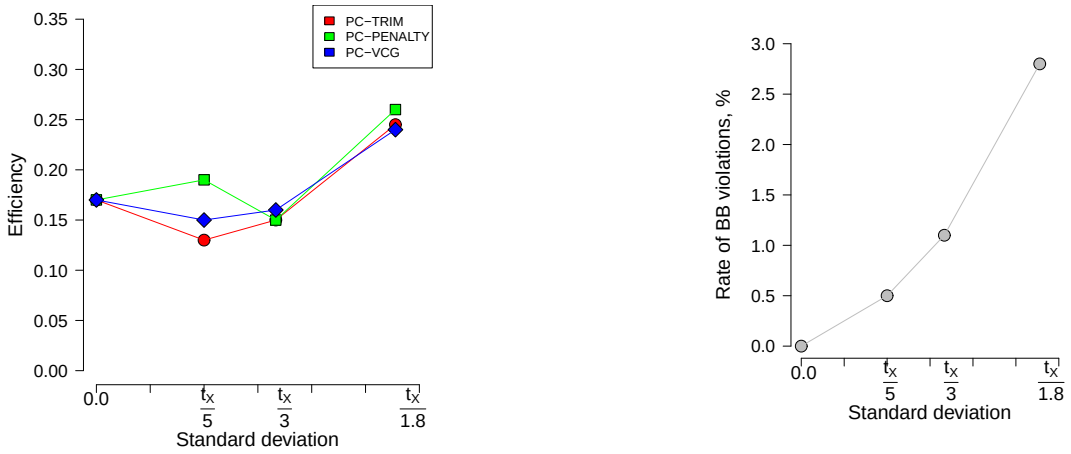

Figure 5: Efficiency and the rate of budget balance violations for different degrees of uncertainty in the scenario with a single buyer and three sellers

point Δ^{trg} in the domain with uncertainty or by the fact that in an uncertain domain it is easy for an agent to lose some of its utility by manipulating because of a more unpredictable market behavior. To check if a higher uncertainty can really increase the efficiency of the market we perform a set of experiments with different levels of uncertainty injected into the market.

To manipulate the degree of uncertainty of our market we vary the σ parameter at the error injection step. Remember, that the error is injected in the following way $t_X^{alc} \sim \mathcal{N}(t_X^{est}, \sigma)$, where t_X^{est} is the estimated number of triples reported by a seller s_X, $X \in \{A, B, C, ...\}$ (see Section 5.2). Thus, by increasing σ we increase the uncertainty in the domain. In all our experiments we parametrize $\sigma = \frac{t_X^{est}}{\alpha}$, where $\alpha \in \mathbb{R}_+$ is a parameter used to manipulate the uncertainty level. We perform a number of experiments with $\alpha = 1.8$, $\alpha = 3$, $\alpha = 5$, and $\alpha \to \infty$ (which corresponds to a domain without uncertainty). Note that in the case with $\alpha \to \infty$ (and, consequently, $\sigma = 0$) all the proposed payment rules are essentially equivalent to the standard *Threshold* rule.

Figures 5-7 illustrate the dependency of the efficiency and the rate of budget balance violations on the uncertainty level of the domain. For all considered scenarios with different number of agents we see a positive trend in efficiency when σ increases. In the case with three sellers there is a small drop in efficiency when increasing σ from 0 to $\frac{t_X}{3}$. We explain this by the fact that in the domain with very influential (or monopolistic) sellers agents can still manipulate a lot even despite of a small uncertainty in the domain. However, when σ is large enough it becomes difficult to manipulate even for influential agents and thus we again have a positive trend in efficiency when $\sigma > \frac{t_X}{3}$.

It is misleading, however, to think that we can always increase the efficiency of the market by injecting more uncertainty in the domain. The problem which arises here is that when we increase the uncertainty in the domain we also increase the number of budget balance violations (see the right-side Figures 5-7). This is not surprising as larger differences between t^{est} and t^{als} (due to larger σ) lead to a higher amounts of inefficient allocations with a total cost of a plan larger than the value of a buyer for the plan. Thus, there is a trade-off between a high efficiency and a budget balance property. Important also that for the large number of sellers the trend of

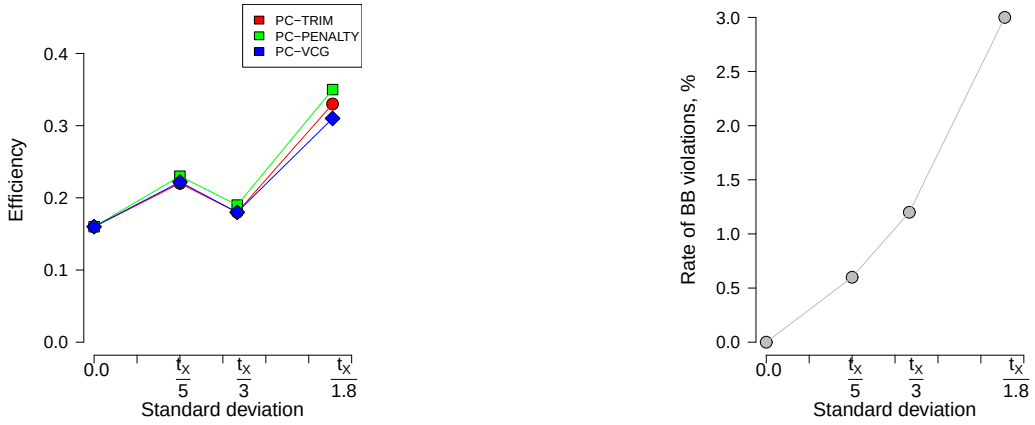

Figure 6: Efficiency and the rate of budget balance violations for different degrees of uncertainty in the scenario with a single buyer and five sellers

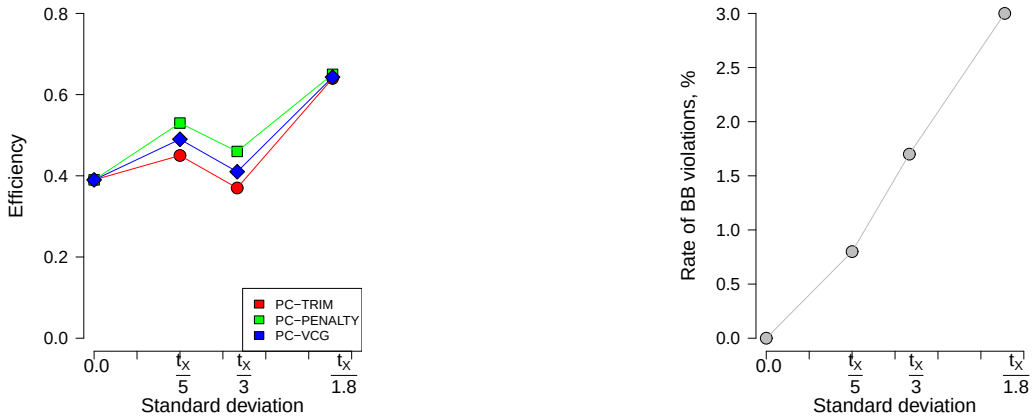

Figure 7: Efficiency and the rate of budget balance violations for different degrees of uncertainty in the scenario with a single buyer and 10 sellers

budget balance violations is almost linear. This means that small perturbations of the uncertainty in statistics provided by sellers does not lead to a dramatic growth of the budget required to cover all inefficiencies occurred due to these imprecise estimates.

6 Conclusion

In this paper, we have proposed a double auction for querying the WoD which handles the fact that the input to both, the allocation and pricing rules, are uncertain due to statistical estimation errors. We have proposed a general market framework for this domain, and introduced a formal model for the double auction. It turns out that a particular challenge in our domain was to design a pricing rule that satisfies the participation constraint despite uncertainty. In particular, the *Treshold* rule, which is guaranteed to satisfy participation in standard double auction environments, can now sometimes lead to an infeasible pricing problem. Towards this end, we have proposed three payment correction rules which always guarantee the participation constraint.

We have evaluated the three rules via a computational Bayes-Nash equilibrium analysis. Interestingly, we have found that incentives, the efficiency, and profits of our new rules are relatively competitive with the *Threshold* rule in a uncertainty-free setting, despite the significant influence of imprecise estimates on the market. We have also evaluated the impact of the number of sellers (and their influence), and found that, consistent with intuition, a larger number of sellers leads to higher efficiency, less manipulation, and higher profits. Furthermore, we have studied the amount of budget required to cover the budget deficit which still occurs, even under our price correction rules, and found this to be almost negligible. Overall, our evaluation results look promising toward implementing an auction-based market for the WoD.

In future work, we are planning to investigate how relaxing some of our assumptions (like those regarding the value and cost functions of the agents) may affect the efficiency, profits and incentive properties of the market. Furthermore, we are going to investigate the impact of a more realistic error injection module, by incorporating a real Query Planner as well as real data stores into our simulation setup.

References

Anandalingam, G., Robert W. Day, and S. Raghavan. 2005. "Landscape of Electronic Market Design." *Management Science*, 51(3): 316–327.

Antoniou, Grigoris, and Frank van Harmelen. 2004. *A Semantic Web Primer.* Cambridge, Massachusetts, London, England: The MIT Press.

Ausubel, Lawrence M., and Oleg V. Baranov. 2010. "Core-Selecting Auctions with Incomplete Information."

Auyoung, Alvin, Laura Grit, Janet Wiener, and John Wilkes. 2006. "Service Contracts and Aggregate Utility Functions." In *In Proceedings of the IEEE Symposium on High Performance Distributed Computing.*

Cramton, Peter. 2013. "Spectrum Auction Design." *Review of Industrial Organization*, 42(2): 030–190.

Cyganiak, Richard, David Wood, and Markus Lanthaler. 2014. "RDF 1.1 Concepts and Abstract Syntax." *http://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/*.

Dash, Debabrata, Verena Kantere, and Anastasia Ailamaki. 2009. "An Economic Model for Self-tuned Cloud Caching." In *Proceedings of the 2009 IEEE International Conference on Data Engineering.*

Erling, O, and I Mikhailov. 2009. "RDF Support in the Virtuoso DBMS." OpenLink Software.

Goeree, Jacob, and Yuanchuan Lien. 2014. "On the Impossibility of Core-Selecting Auctions." *Theoretical Economics*. Forthcoming.

Goetzendorf, Andor, Martin Bichler, Pasha Shabalin, and Robert W. Day. 2015. "Compact Bid Languages and Core Pricing in Large Multi-item Auctions." In *Management Science*.

Görlitz, Olaf, and Steffen Staab. 2011. "SPLENDID: SPARQL Endpoint Federation Exploiting VOID Descriptions." In *Proceedings of the 2nd International Workshop on Consuming Linked Data*. Bonn, Germany.

Harris, Steve, and Andy Seaborne. 2013. "SPARQL 1.1 Query Language." *http://www.w3.org/TR/2013/REC-sparql11-query-20130321/*.

Harth, A, J Umbrich, A Hogan, and S Decker. 2007. "Yars2: a Federated Repository for Querying Graph Structured Data from the Web." *6th International Semantic Web Conference (ISWC)*, 211–224.

Hartig, O, C Bizer, and J C Freytag. 2009. "Executing SPARQL Queries over the Web of Linked Data." *8th International Semantic Web Conference ISWC2009*, 293–309.

Koutris, Paraschos, Prasang Upadhyaya, Magdalena Balazinska, Bill Howe, and Dan Suciu. 2013. "Toward Practical Query Pricing with QueryMarket." In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*.

Labrinidis, Alexandros, Huiming Qu, and Jie Xu. 2007. "Quality Contracts for Real-time Enterprises." In *Proceedings of the 1st international conference on Business intelligence for the real-time enterprises*. Berlin, Heidelberg:Springer-Verlag.

Lai, Kevin, Lars Rasmusson, Eytan Adar, Li Zhang, and Bernardo A. Huberman. 2005. "Tycoon: An Implementation of a Distributed, Market-based Resource Allocation System." *Multiagent and Grid Systems*, 1(3): 169–182.

Langegger, A, W Wöß, and M Blöchl. 2008. "A Semantic Web Middleware for Virtual Data Integration on the Web." *Lecture Notes In Computer Science*.

Lubin, Benjamin, and David C. Parkes. 2009. "Quantifying the Strategyproofness of Mechanisms via Metrics on Payoff Distributions." In *Proceedings of the 17th National Conference on Artificial Intelligence*.

Lubin, Benjamin, Benedikt Bünz, and Sven Seuken. 2015. "Fairness Beyond the Core: New Payment Rules for Combinatorial Auctions." Working Paper.

Malone, Thomas W., Richard E. Fikes, and Michael T. Howard. 1983. "Enterprise : a Market-like Task Scheduler for Distributed Computing Environments."

Milgrom, Paul. 2007. "Package Auctions and Exchanges." *Econometrica*, 75(4): 935–965.

Oren, E, C Gueret, and S Schlobach. 2008. "Anytime Query Answering in RDF through Evolutionary Algorithms." *The Semantic Web - ISWC 2008*, 5318: 98–113.

Parkes, David C., Jayant Kalagnanam, and Marta Eso. 2001. "Achieving Budget-balance with Vickrey-based Payment Schemes in Exchanges." In *Proceedings of the 17th International Joint Conference on Artificial Intelligence*. San Francisco, CA.

Quilitz, B, and U Leser. 2008. "Querying Distributed RDF Data Sources with SPARQL." *Proceedings of the 5th European semantic web conference on The semantic web: research and applications*, 524–538.

Schwarte, Andreas, Peter Haase, Katja Hose, Ralf Schenkel, and Michael Schmidt. 2011. "FedX: Optimization Techniques for Federated Query Processing on Linked Data." In *International Semantic Web Conference (1)*.

Stonebraker, Michael, Paul M. Aoki, Witold Litwin, Avi Pfeffer, Adam Sah, Jeff Sidell, Carl Staelin, and Andrew Yu. 1996. "Mariposa: A Wide-Area Distributed Database System." *VLDB J.*, 5(1): 48–63.

Sutherland, I. E. 1968. "A Futures Market in Computer Time." *Commun. ACM*, 11(6): 449–451.

Van Alstyne, Marshall, Erik Brynjolfsson, and Stuart Madnick. 1995. "Why Not One Big Database? Principles for Data Ownership." *Decis. Support Syst.*, 15(4): 267–284.

Waldspurger, C.A., T. Hogg, B.A. Huberman, J.O. Kephart, and W.S. Stornetta. 1992. "Spawn: a Distributed Computational Economy." *IEEE Transactions on Software Engineering*, 18(2): 103–117.

Zemánek, J, S Schenk, and V Svatek. 2007. "Optimizing SPARQL Queries Over Disparate RDF Data Sources Through Distributed Semi-joins." *7th International Semantic Web Conference (ISWC)*.

2.1 Addendum

Observe, that the setting studied in this paper is not a general double auction setting but a very restricted one. In particular, it includes only a single buyer and multiple sellers. Consequently, this makes our domain look similar to the domain of the standard reverse combinatorial auction. However, in our work, we assume that both the buyer and the sellers can behave strategically (which is not the case in the standard setting of combinatorial auctions). This means that the impossibility result of the Myerson-Satterthwaite theorem applies to our setting, and thus, even the simple VCG mechanism in the domain without uncertainty may lead to a significant rate of budget balance violations.

3 Core-Selecting Payment Rules for Combinatorial Auctions with Uncertain Availability of Goods

> People of the same trade seldom meet together, even for merriment and diversion, but the conversation ends in a conspiracy against the public, or in some contrivance to raise prices.
>
> Adam Smith, The Wealth of Nations

The content of this chapter has previously appeared in:

> Moor, D. and Seuken, S. and Grubenmann, T. and Bernstein, A. (2016). Core-Selecting Payment Rules for Combinatorial Auctions with Uncertain Availability of Goods. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence*, IJCAI '16, New York, NY, USA.

Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16)

Core-Selecting Payment Rules for Combinatorial Auctions with Uncertain Availability of Goods

Dmitry Moor and **Sven Seuken** and **Tobias Grubenmann** and **Abraham Bernstein**
Department of Informatics
University of Zurich
Zurich, Switzerland
{dmitry.moor, seuken, grubenmann, bernstein}@ifi.uzh.ch

Abstract

In some auction domains, there is uncertainty regarding the final availability of the goods being auctioned off. For example, a government may auction off spectrum from its public safety network, but it may need this spectrum back in times of emergency. In such a domain, standard combinatorial auctions perform poorly because they lead to violations of individual rationality (IR), even in expectation, and to very low efficiency. In this paper, we study the design of core-selecting payment rules for such domains. Surprisingly, we show that in this new domain, there does not exist a payment rule with is guaranteed to be ex-post core-selecting. However, we show that by designing rules that are "execution-contingent," i.e., by charging payments that are conditioned on the realization of the availability of the goods, we can reduce IR violations. We design two core-selecting rules that always satisfy IR in expectation. To study the performance of our rules we perform a computational Bayes-Nash equilibrium analysis. We show that, in equilibrium, our new rules have better incentives, higher efficiency, and a lower rate of ex-post IR violations than standard core-selecting rules.

1 Introduction

Combinatorial auctions (CAs) have been successfully applied in many real-world settings, including procurement auctions [Sandholm, 2013], TV advertising auctions [Goetzendorf *et al.*, 2015], and government spectrum auctions [Cramton, 2013; Ausubel and Baranov, 2014]. CAs are specifically designed for domains where bidders can have complex preferences over bundles of heterogeneous items. An important and seemingly "innocent" assumption in auction design is that all of the goods will be available for consumption by the winning bidders. While this assumption is often satisfied, there are some domains where this is not the case. In these domains, standard mechanisms perform poorly and new designs are required.

1.1 Uncertain Availability: A Motivating Example

In the US, public safety networks are used by the police, firefighters, and emergency medical technicians during times of emergencies. In 2012, following the events of 9/11 and hurricane Katrina, the US congress even reserved some parts of the 700MHz spectrum "to public safety for use in a nationwide broadband network" [FCC, 2016]. However, this legislation also allows for the use of this spectrum by private companies when the spectrum would otherwise be idle. The company *Rivada Networks* for example is currently designing an auction platform for this purpose [Cramton and Doyle, 2015]. These auctions can happen on a weekly or daily basis, or even in real-time; but bidders have to accept the risk that the spectrum they purchased in the auction might become unavailable because an emergency occured and the spectrum is needed for public safety reasons.

There are many other CA domains with this kind of *uncertain availability of goods*. For example, the company *BandwidthX* has developed a platform to auction off bandwidth from wireless hotspots, but those hotspots can also become unavailable at any point in time [Seuken *et al.*, 2015]. Furthermore, Moor et al. [2015] introduced an auction for data, where the sellers only have imprecise estimates of the data they will actually be able to supply.

1.2 Execution-Contingent Auctions

Domains with uncertain availability of goods pose new challenges. As we will show, standard mechanisms perform very poorly: they violate individual rationality (IR), even in expectation, they have bad incentives, and low efficiency in equilibrium. To address these shortcomings, Porter et al. [2008] introduced *execution-contingent* mechanisms, which charge payments dependent on the realized availabilities of the goods (see [Ceppi *et al.*, 2015; Ramchurn *et al.*, 2009] for extensions). Using this paradigm, one can design mechanisms that satisfy IR and strategyproofness *in expectation*. For domains as described above, where many small auctions may be run repeatedly over time, *some* IR violations may be acceptable, as long as IR in expectation is satisfied.

However, the execution-contingent mechanisms studied in the literature so far [Porter *et al.*, 2008; Ceppi *et al.*, 2015; Ramchurn *et al.*, 2009] are not suitable for CAs in practice because they charge VCG-like payments. Unfortunately, in a CA domain, VCG has numerous problems [Ausubel and Milgrom, 2006]. Most notably, VCG can lead to very low or even zero revenue, which opens up opportunities for collusion between the seller and collections of bidders.

1.3 Core-Selecting Payment Rules

These drawbacks of VCG have led to the development of *core-selecting payment rules* which offer a principled way to ensure that the revenue in the auction is high enough such that there are no opportunities for collusion [Milgrom, 2007; Day and Milgrom, 2008; Day and Raghavan, 2007; Ausubel and Baranov, 2016]. They have already been used successfully in practice, for example in the UK, Canada, the Netherlands, and Switzerland to auction off billions of dollars worth of 4G spectrum. In this paper, we design *execution-contingent core-selecting payment rules* for domains with uncertainty availability of goods. The first question we ask is "what is the right notion of an execution-contingent core?" Second, we find that our execution-contingent cores may sometimes be empty, and thus, our rules will only be "core-selecting" in a relaxed sense. Finally, in contrast to prior work on execution-contingent mechanisms, we are not fully content with rules that satisfy "IR in expectation," but also want to minimize the rate of ex-post IR violations.

1.4 Overview of Contributions

Our goal in this paper is to design payment rules that work well in a domain with uncertain availability of goods. We make the following contributions:

1. We generalize the EC-VCG mechanism introduced by Porter et al. [2008] to domains with *continuous* and *dependent* availabilities of goods.
2. We introduce two execution-contingent core-selecting payment rules which satisfy IR in expectation.
3. We perform a computational Bayes-Nash equilibrium (BNE) analysis evaluating our new rules in terms of incentives, efficiency, and ex-post IR violations.

The results from our BNE analysis show that our new payment rules have better incentives, higher efficiency, and a lower rate of ex-post IR violations than standard core-selecting rules.

2 Preliminaries

2.1 Formal Model

Let $N = \{1, ..., n\}$ be a set of bidders and let s be the seller. Let $G = \{A, B, C, ...\}$ be the set of goods, with $|G| = m$. Each bidder has a valuation function $v_i : 2^G \to \mathbb{R}$ which specifies i's value for every possible bundle $S \in 2^G$.

We consider a *two time-period* model: at *allocation time*, there is uncertainty regarding which bundles will eventually be available, and at *consumption time*, some of the availability (depending on the mechanism) will be revealed. With every bundle S we associate a random variable $a(S)$ which represents the *availability* of the bundle S, i.e., the extent to which S will be available at the time of consumption. We let $\tilde{a}(S)$ denote the realization of the random variable $a(S)$ at consumption time.

Let f be a joint probability mass function for all random variables $a(S)$. We assume that f is exogenous and known by all agents.

The seller may have costs for providing the goods. We denote costs by $\{c_1, ..., c_m\}$, where c_j is the cost of the j-th good. We assume that the cost function $c(S) \in \mathbb{R}$ is additive, such that $c(S) = \sum_S c_j$. We assume that the seller does not strategize, as is common in the analysis of CAs [Day and Cramton, 2012].

An allocation is denoted as $x = \{x_1, ..., x_n\}$, with x_i denoting the bundle allocated to agent i and x_{-i} denoting the vector of bundles of all agents except i. Similarly, for any $K \subset N$, we use x_K when referring to an allocation among all bidders in K and use x_{-K} to denote an allocation among bidders in $N \backslash K$. We let $v_i(x_i)$ denote bidder i's true value for its allocated bundle x_i, and we let $\hat{v}_i(x_i)$ denote bidder i's value report for bundle x_i (possibly non-truthful). We let $a(x_i)$ denote the random variable corresponding to bundle x_i, and $\tilde{a}(x_i^*)$ denote the *realized availability* corresponding to x_i^*. If bidder i is allocated bundle x_i^* then his *realized value* is thus given by $v_i(x_i^*)\tilde{a}(x_i^*)$.

Assumption 1. *We assume that all bidders are* "extremely" single-minded*, i.e., each bidder has non-zero value for exactly one bundle $S \subseteq 2^G$, and zero value for all other bundles $S' \neq S$ (including supersets of S).*

Remark 1. *This assumption allows for the simple definition of the realized value we have just provided, i.e., $v_i(x_i^*)\tilde{a}(x_i^*)$. Without this assumption, we would have to consider all sub-bundles of a bidder's allocated bundle to compute his realized value, i.e., $\max_{S \subseteq x_i^*}\{v(S)\tilde{a}(S)\}$, which may be computationally infeasible in domains with a large number of goods. But more importantly, we make this assumption to simplify the notation and the analysis of the mechanisms we will present. In future work, we will extend our results to the full domain.*

Let x be an allocation. We let W_x (the *winners*) denote the set of allocated agents under this allocation, i.e., $W_x = \{i | x_i \neq \emptyset\}$. The social welfare of the allocation x is

$$SW(x) = \sum_{i \in W_x} (v_i(x_i) - c(x_i))\tilde{a}(x_i).$$

We assume quasilinear utilities $u_i(x_i, p_i) = v_i(x_i) - p_i$, where p_i is bidder i's payment for x_i; we let p_s denote the payment received by the seller. We let $p = (p_s, p_1, p_2, ..., p_n)$ denote the vector of payments received by the seller and paid by all bidders. We let $O = \langle x, p \rangle$ denote an outcome, i.e., an allocation and the payment vector.

2.2 Properties of Mechanisms

Let $v = (v_1, ..., v_n)$ and $c = (c_1, ..., c_m)$. We let $\tilde{a}$ denote the vector of availabilities at consumption time (which may depend on the mechanism and the domain), i.e., $\tilde{a} = (\tilde{a}(S) : \forall S \in 2^G, \tilde{a}(S)$ is known at consumption time). We let $\mathcal{M} = \langle g, h \rangle$ denote a mechanism, where $g(v, c, f) = x$ is an allocation rule and $h(x, v, c, f, \tilde{a}_M) = p$ is a payment rule. We now define a number of standard mechanism design properties; however, because we consider a domain with uncertainty, we need to define most of these properties "in expectation."

Definition 1. *A mechanism $\mathcal{M} = \langle g, h \rangle$ is **strategyproof in expectation** if $\forall i \in N$, $\forall v_i$, for all $\hat{v}_{-i}$*

$$\mathbb{E}_f[u_i(g(v_i, \hat{v}_{-i}, c, f), p_i)] \geq \mathbb{E}_f[u_i(g(\hat{v}_i, \hat{v}_{-i}, c, f), p_i)].$$

Definition 2. *A mechanism $\mathcal{M} = \langle g, h \rangle$ with $h(g(v, c, f), v, c, f, \tilde{a}_M) = p$ is ex-post **individually rational (IR)** if $\forall i \in N$, $\forall v_i, \forall \hat{v}_{-i}$*

$$v_i(g(v_i, \hat{v}_{-i}, c, f))\tilde{a}(g(v_i, \hat{v}_{-i}, c, f)) - p_i \geq 0.$$

Definition 3. *A mechanism $\mathcal{M} = \langle g, h \rangle$ with $h(g(v, c, f), v, c, f, \tilde{a}_M) = p$ is **individually rational in expectation (IRE)** if $\forall i \in N$, $\forall v_i$, for all $\hat{v}_{-i}$*

$$\mathbb{E}_f[u_i(g(v_i, \hat{v}_{-i}, c, f), p_i)] \geq 0.$$

Definition 4. *The **rate of ex-post IR violations** of a mechanism $\mathcal{M} = \langle g, h \rangle$ with $h(g(v, c, f), v, c, f, \tilde{a}_M) = p$ is defined as the following probability:*

$$\mathbb{P}(v_i(g(v_i, \hat{v}_{-i}, c, f))\tilde{a}(g(v_i, \hat{v}_{-i}, c, f)) - p_i < 0) \quad (1)$$

Definition 5. *A mechanism is **budget balanced** if the sum of all payments paid by the bidders is equal to the payment received by the seller, i.e.,*

$$\sum_{i \in N} p_i = p_s.$$

Definition 6. *A mechanism is **expected social welfare maximizing** if its allocation rule selects an allocation x with*

$$x \in \operatorname*{argmax}_x \mathbb{E}_f[SW(x)].$$

2.3 VCG Mechanism

The famous VCG mechanism [Vickrey, 1961; Clarke, 1971; Groves, 1973] selects a social welfare maximizing allocation and computes payments equal to the externality each agent imposes on all other agents. We let x^* denote the allocation which maximizes social welfare when all agents are considered, and x^{-i} denotes the allocation which maximizes social social welfare when all agents except i are considered. VCG payments are then:

$$p_i^{\text{VCG}} = SW(x^{-i}) - SW_{-i}(x^*).$$

VCG is a particularly attractive mechanism because, in a domain without uncertainty about the availability of goods, it is social welfare maximizing, strategyproof and satisfies IR.

3 Execution-Contingent VCG

Porter et al. [2008] generalized the VCG mechanism to domains with uncertain availability of goods by introducing an *execution-contingent* variant of VCG. The main idea is to make payments contingent on the realized availabilities, which implies that the payments are not computed at allocation time, but at consumption time (see Figure 1). They considered a domain with binary and independent random variables capturing the availabilities, and proved that in this domain, their mechanism is strategyproof and IR in expectation. However, in the domains we described in the beginning, the availabilities of the goods will typically be *dependent* (e.g., consider a terrorist attack affecting a whole city) and the availabilities may be continuous (e.g., a resource can be used partially). We now introduce the ECC-VCG mechanism, which generalizes the mechanism introduced by Porter et al. [2008] to also handle continuous, dependent random variables.

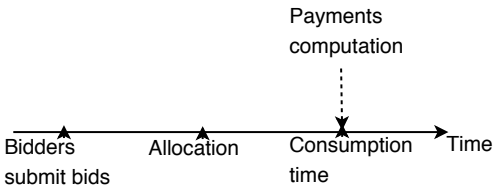

Figure 1: Flow chart of execution-contingent mechanisms.

ECC-VCG Mechanism.

- *Allocation rule*: select $x^* \in \operatorname{argmax}_x \mathbb{E}_f[SW(x)]$
- *Payment rule*:

$$p_i^{\text{ECC-VCG}} = \mathbb{E}_f[SW(x^{-i})|\tilde{a}(x_j^*), j \in W_{x^*}] - SW_{-i}(x^*)$$

The idea behind this *execution-contingent conditional VCG* mechanism is similar to standard VCG: $p_i^{\text{ECC-VCG}}$ represents a bidder's externality on all other bidders, except that we do not know this externality exactly and thus compute an estimate by taking the conditional expectation. The following example illustrates the ECC-VCG mechanism.

Example 1. *Consider a setting with three bidders $N = \{1, 2, 3\}$, two goods $G = \{A, B\}$ and a seller s with no costs. The bidders' values are provided in the following table:*

	A	B	$\{A, B\}$
Bidder 1	*0.1*		
Bidder 2		*0.2*	
Bidder 3			*0.3*

Let $a(\{A\}), a(\{B\}), a(\{A, B\})$ denote availabilities of $\{A\}, \{B\}$ and $\{A, B\}$ respectively; let f denote the joint probability mass function for these random variables. The following table specifies f:

(a) $a_1 = 0$

a_2	$a_3 = 0$	$a_3 = 1$
0	0.25	0
1	0.25	0

(b) $a_1 = 1$

a_2	$a_3 = 0$	$a_3 = 1$
0	0.1	0
1	0	0.4

It is easy verify that the expected marginal availabilities of the bundles $\{A\}$, $\{B\}$, and $\{A, B\}$ are 0.5, 0.65 and 0.4 respectively. Thus, the efficient allocation is to allocate A to bidder 1 and B to bidder 2, and $\mathbb{E}_f[SW(x)] = 0.1 \cdot 0.5 + 0.2 \cdot 0.65 = 0.18$.

Now assume that after the allocation has happened the realized availabilities are $\tilde{a}_1 = 0$ and $\tilde{a}_2 = 1$. Thus $\mathbb{E}_f[a_3|\tilde{a}_1, \tilde{a}_2] = 0$. The execution-contingent payments are:

$$p_1^{ECC-VCG} = 0.2 \cdot \tilde{a}_2 - (0.2 \cdot \tilde{a}_2) = 0 \quad (2)$$

$$p_2^{ECC-VCG} = 0 - (0.1 \cdot \tilde{a}_1) = 0 \quad (3)$$

However, if realized availabilities were $\tilde{a}_1 = 1$ and $\tilde{a}_2 = 1$, then $\mathbb{E}_f[a_3|\tilde{a}_1, \tilde{a}_2] = 0.4$ and execution contingent payments would be

$$p_1^{ECC-VCG} = 0.2 \cdot \tilde{a}_2 - (0.2 \cdot \tilde{a}_2) = 0 \quad (4)$$

$$p_2^{ECC-VCG} = 0.3 \cdot \mathbb{E}_f[a_3|\tilde{a}_1, \tilde{a}_2] - (0.1 \cdot \tilde{a}_1) = 0.02. \quad (5)$$

Note that the ECC-VCG mechanism only takes the realization of the *allocated* bundles into account when computing payments. If we know the realizations of *all* bundles at consumption time, then we can use the following mechanism:

ECR-VCG Mechanism.

- *Allocation rule*: select $x^* \in \text{argmax}_x \, \mathbb{E}_f[SW(x)]$
- *Payment rule*:

$$p_i^{\text{ECR-VCG}} = \mathbb{E}_f[SW(x^{-i})|\tilde{a}(S), \forall S] - SW_{-i}(x^*)$$

This *execution-contingent realized VCG* mechanism is a special case of ECC-VCG, as the payment rule uses a more precise estimate of the externality imposed on other bidders.

Theorem 1. *ECC-VCG and ECR-VCG are strategyproof in expectation.*

Proof. We prove the theorem for ECC-VCG; the proof for ECR-VCG is analogous. We let $\hat{v}_i$ denote a reported value function of agent $i \in N$ and v_i is the true value of the agent. Let x^* be the optimal allocation given $(\hat{v}_i, \hat{v}_{-i})$. Consider the conditional expectation in the payment rule:

$$\mathbb{E}_f[SW(x^{-i})|\tilde{a}(x_j^*), j \in W_{x^*}] \tag{6}$$

$$= \sum_{k \in W_{x^{-i}}} (\hat{v}_k(x_k^{-i}) - c(x_k^{-i}))\mathbb{E}_f[a(x_k^{-i})|\tilde{a}(x_j^*), j \in W_{x^*}]. \tag{7}$$

Now the payment of a bidder $i \in W_{x^*}$ is

$$p_i^{\text{ECC-VCG}} = \sum_{k \in W_{x^{-i}}} (\hat{v}_k(x_k^{-i}) - c(x_k^{-i}))\mathbb{E}_f[a(x_k^{-i})|\tilde{a}(x_j^*), j \in W_{x^*}]$$

$$- \sum_{k \in W_{x^*} \setminus i} (\hat{v}_k(x_k^*) - c(x_k^*))\tilde{a}(x_k^*) + c(x_i^*)\tilde{a}(x_i^*)$$

The expected utility of bidder i in this case is

$$\mathbb{E}_f[u_i(x^*)] = \mathbb{E}_f[v_i(x_i^*)a(x_i^*) - p_i^{\text{ECC-VCG}}] =$$

$$\mathbb{E}_f\Big[v_i(x_i^*)a(x_i^*) - c(x_i^*)a(x_i^*) + \sum_{k \in W_{x^*} \setminus i} (\hat{v}_k(x_k^*) - c(x_k^*))a(x_k^*)$$

$$\underbrace{- \sum_{k \in W_{x^{-i}}} (\hat{v}_k(x_k^{-i}) - c(x_k^{-i}))\mathbb{E}_f[a(x_k^{-i})|a(x_j^*), j \in W_{x^*}]\Big]}_{\texttt{does not depend on } \hat{v}_i}$$

The first three terms under the last expectation is equal to expected social welfare if the agent reports truthfully. Thus, because the allocation rule maximizes expected social welfare, and because the last expression is independent of i's report, truthful reporting maximizes agent i's expected utility. In this case: $\mathbb{E}_f[u_i(x^*)] = \mathbb{E}_f[SW(x^*)] - \mathbb{E}_f[SW(x^{-i})]$. □

Theorem 2. *ECC-VCG and ECR-VCG are individually rational in expectation.*

Proof. We show the theorem for ECC-VCG. From the proof of Theorem 1 we know that under truthful reporting

$$\mathbb{E}_f[u_i(x^*)] = \mathbb{E}_f[SW(x^*)] - \mathbb{E}_f[SW(x^{-i})]. \tag{8}$$

Because the allocation rule selects an expected social welfare maximizing allocation we know that $\mathbb{E}_f[SW(x^*)] \geq \mathbb{E}_f[SW(x^{-i})])$, and thus $\mathbb{E}_f[u_i(x^*)] \geq 0$. The proof is analogous for ECR-VCG. □

4 Core-Selecting Payment Rules

VCG-based rules are attractive because they are strategyproof (or strategyproof in expectation, in our domain). However, they are not suitable in CAs because they can lead to very low or even zero revenue, which opens up opportunities for collusion between the seller and collections of bidders. In this section, we introduce our *execution-contingent core-selecting payment rules*. To this end, we first need some definitions:

Definition 7. *An outcome O is blocked by a coalition $K \subset N$ of bidders, if there exists another outcome $\bar{O}$ which is weakly preferred over O by every bidder $i \in K$ and which provides higher utility for the seller. In this case K is called a blocking coalition.*

Definition 8. *An outcome O is in the core if it is (a) individually rational and (b) is not blocked by any coalition.*

If $O^* = \langle x^*, p \rangle$ and $K \subset N$ is a coalition of bidders, then $\sum_{i \in K}(v_i(x_i^*) - p_i)$ is the total opportunity cost of agents in the coalition. We let $\omega(K)$ denote the total welfare which the coalition could achieve by redistributing goods among themselves. The maximum additional value which K could give to the seller is $\omega(K) - \sum_{i \in K}(v_i(x_i^*) - p_i)$. A core constraint guarantees that this amount is not larger than what the seller can already get under the current allocation, and this core constraint has to hold for all possible coalitions:

$$\sum_{i \in W_{x^*}} (p_i - c(x_i^*)) \geq \omega(K) - \sum_{i \in K}(v_i(x_i^*) - p_i) \quad \forall K \subseteq N$$

A core-selecting mechanism selects a social welfare maximizing allocation and picks payments from the core. Unfortunately, in general combinatorial auction domains (with complements), even without uncertain availability of goods, there does not exist a payment rule that is Bayes-Nash incentive compatible and core-selecting [Goeree and Lien, 2016]. Of course, this impossibility extends to our new domain. Thus, none of our core-selecting payment rules can be truthful.

The core in a domain with uncertainty. In contrast to the standard domain, defining the core in a domain with uncertain availability of goods is less straightforward. The difficulty arises from the fact that in a domain with uncertainty agents might be willing to deviate either before or after availabilities of goods are realized. In the former case, we can talk about an *ex-ante* core while the latter case implies an *ex-post* core. Although these two concepts differ only in the amount of information which is used to evaluate the expected total welfare achieved by a coalition, we will show that these concepts actually lead to quite different properties.

In the next section, we will first present an ex-ante core-selecting payment rule, which is not execution contingent. However, as we will later see in Section 5, this leads to a relatively large rate of IR violations, compared to execution-contingent payment rules. To address this, we will then study ex-post core-selecting payment rules in Sections 4.2-4.5.

4.1 Ex-ante Core-Selecting Payment rules

A straightforward approach to generalize the idea of a core to a domain with uncertainty is to apply the core constraints

as defined in the previous section to the expected values of the bidders. In this case the total expected welfare $\mathbb{E}_f[\omega(K)]$ that a coalition K can achieve in expectation is $\mathbb{E}_f[\omega(K)] = \max_x \mathbb{E}_f[\sum_{i\in K}(v_i(x) - c_i(x))a(x)]$. By limiting this value to be smaller than the expected total utility which agents in the coalition can get under the current allocation we ensure that the agents in the coalition are not be willing to deviate from the current allocation. Formally, this gives rise to the following set of core constraints, which have to hold for all $K \subseteq N$:

$$\mathbb{E}_f\Big[\sum_{i\in W_{x^*}}(p_i - c(x_i^*)a(x_i^*))\Big] \geq \tag{9}$$

$$\mathbb{E}_f\Big[\omega(K)\Big] - \sum_{i\in K}\mathbb{E}_f\Big[(v_i(x_i^*)a(x_i^*) - p_i)\Big] \tag{10}$$

Even though this core is not execution contingent, it provides all core properties ex-ante, i.e., before availabilities of goods are realized. Furthermore, the following corollary says that this core is never empty:

Corollary 1. *In a domain with uncertain availability of goods, the ex-ante core is never empty.*

The corollary follows from the fact that the core must contain at least one point, namely pay-as-bid (see [Day and Milgrom, 2008, Footnote 1]).

4.2 Impossibility of Ex-post Core-Selecting Payment Rules in Domains with Uncertainty

In this section, we will show the surprising result that there does not, in general, exist an ex-post core-selecting payment rule in a domain with uncertain availability of goods. For this, we first need a few more definitions. Assume that $\langle x^*, p\rangle$ is an auction outcome and let $L \subseteq 2^G$ be a fixed set of bundles.

Definition 9. *A generalized expected coalitional value $\omega_{ge}(K, L)$ of a coalition $K \subseteq N$ given the set L is the maximum expected welfare the coalition can achieve given realized availabilities corresponding to bundles from L. Formally,*

$$\omega_{ge}(K, L) = \max_x \mathbb{E}_f\Big[\sum_{i\in K}(v_i(x_i) - c(x_i))a(x_i)\Big|\tilde{a}(S), S \in L\Big]$$

Note that bidders have a total opportunity cost of $\sum_{i\in K}(v_i(x_i^*)\tilde{a}(x_i^*) - p_i)$ for joining coalition K. If they decide to join, then in expectation they can achieve a total welfare of $\omega_{ge}(K, L)$ and thus they can provide at most $\omega_{ge}(K, L) - \sum_{i\in K}(v_i(x_i^*)\tilde{a}(x_i^*) - p_i)$ of additional value to the seller, which gives rise to the following set of core constraints:

Generalized Ex-post Core Constraint $\forall K \subseteq N$:

$$\sum_{i\in W_{x^*}}(p_i - c(x_i^*)\tilde{a}(x_i^*)) \geq \omega_{ge}(K, L) - \sum_{i\in K}(v_i(x_i^*)\tilde{a}(x_i^*) - p_i)$$

Note that different choices of L lead to a different definitions of an ex-post core. However, we refer to any of these cores as *ex-post* cores, which reflects the fact that the core property is guaranteed for the point in time when availabilities have been realized.

Unfortunately, in our domain, the ex-post core can sometimes be empty.

Theorem 3. *In a domain with uncertainty, there does not exist a mechanism that is budget balanced and ex-post core-selecting.*

Proof. Consider a setting with two bidders $N = \{1, 2\}$ and a seller s. Assume that there are only two goods A and B. We assume that the first bidder has an value $v_1(\{A\})$ for a bundle $\{A\}$ and the second bidder has a value $v_2(\{AB\})$ for a bundle $\{AB\}$. We let $v_1(\{A\}) > 0$ and $v_2(\{AB\}) > 0$. We also let the costs of those goods to be equal to zero.

Given this specific set-up, there are seven different allocation rules we must consider: the allocation rule can allocate none of the items; it can allocate item A to bidder 1 and nothing to bidder 2; it can allocate item A to bidder 2 and nothing to bidder 1; etc. For each of these seven different allocation rules, there exists a joint probability mass function and a set of realized availabilities, such that we can construct an empty core.

We first consider the case where bidder 1 is allocated item A and bidder 2 is allocated the empty set. We let $a(\{A\})$ and $a(\{AB\})$ denote the availabilities of the bundles $\{A\}$ and $\{AB\}$ respectively, and we let f denote the corresponding joint probability mass function, which is defined as follows:

(a) $a(\{AB\}) = 0$

$a(A)$	$a(B)=0$	$a(B)=1$
0	γ_{000}	γ_{010}
1	γ_{100}	γ_{110}

(b) $a(\{AB\}) = 1$

$a(B)=0$	$a(B)=1$
γ_{001}	γ_{011}
γ_{101}	γ_{111}

Here $0 < \gamma_{ijk} < 1$, $i, j, k \in \{0, 1\}$, denotes the probability of the event that $a(\{A\}) = i$, $a(\{B\}) = j$, and $a(\{AB\}) = k$. Now we assume that $\tilde{a}(\{A\}) = 0$, $\tilde{a}(\{B\}) = 0$, $\tilde{a}(\{AB\}) = 1$. We can show that for any $L \subseteq 2^G$, the corresponding ex-post core is empty. Indeed, $\forall L \subseteq 2^G : \mathbb{E}[a(\{AB\})|\tilde{a}(S), S \in L] > 0$. Thus, the ex-post core constraint is then:

$$\omega_{ge}(\{s, 2\}, L) =$$
$$v_2(\{AB\}) \cdot \mathbb{E}[a(\{AB\})|\tilde{a}(S), S \in L] \leq 0 + p_s = p_1.$$

Here, p_1 and p_s are payments of the first bidder and the seller respectively (which are equal, given budget balance, i.e., $p_1 + p_s = 0$). Now, taking into account that $v_2(\{AB\}) > 0$ but $p_1 \leq v_1(\{A\})\tilde{a}(\{A\}) = 0$ (another core constraint), we get a contradiction. The proof for the other six possible allocations is analogous. □

4.3 Framework for Execution-Contingent Mechanisms

Given that the ex-ante core can lead to large IR violations (as we will show in Section 5), and given the impossibility result regarding ex-post cores from Theorem 3, this raises the question which core to consider in practice. In this section, we put forward the idea of designing mechanisms that select payments inside an ex-post core *whenever this core is not empty*. Specifically, we design two *execution-contingent* payment rules. As we will later show in Section 5, for these payment rules, the empty core cases happen relatively rarely.

We will now first define a *mechanism framework* for execution-contingent payment rules which we then instantiate in two different ways:

Execution-Contingent Core-selecting Mechanism Framework

- *Framework parameters*:
  1. Reference point: p^*
  2. Core constraints: $Core^*$
- *Allocation rule*: select $x^* \in \text{argmax}_x \, \mathbb{E}_f[SW(x)]$
- *Payment rule*:

$$p = \begin{cases} p \in \text{argmin}_{p \in \Pi} ||p - p^*||_2 & \text{if Core*} \cap \text{IR} \neq \emptyset \\ p* & \text{else} \end{cases}$$

where $\Pi = Core^* \cap IR \cap MRC$.

The first framework parameter, p^*, is a reference point, which we will either instantiate to $p^* = p^{\text{ECC-VCG}}$ or $p^{\text{ECR-VCG}}$, as defined by the ECC-VCG and ECR-VCG mechanisms. The second framework parameter is a set of core constraints $Core^*$, which we will accordingly instantiate to ECC-Core or ECR-Core, to be defined next. The ultimate core-selecting payment rule then first tries to find a payment vector p that minimizes the Euclidean distance to the reference point p^*, from among all payment vectors in the core that also minimize the revenue for the seller (the so-called *minimize-revenue constraint (MRC)*). We defined the mechanism framework this way to be analogous to the *Quadratic rule* [Day and Cramton, 2012], i.e., the core-selecting rule most commonly used in practice. In a domain with uncertain availabilities, however, the core (i.e., the intersection of the core constraints and the IR constraints) can be empty. In this case, we charge the reference point p^*, which will then be outside the core. Thus, all of these execution-contingent mechanisms are only ex-post core selecting whenever the core is non-empty. In Section 5, we will analyze how often such "empty core" cases occur in equilibrium.

4.4 ECC-Core Mechanism

Before we can introduce the ECC-Core mechanism, we need one more definition.

Definition 10. *An expected coalitional value $\omega_e(K)$ of a coalition $K \subset N$ is the maximum expected welfare the coalition can achieve knowing realized availabilities of allocated bundles. Formally,*

$$\omega_e(K) = \max_x \mathbb{E}_f\Big[\sum_{i \in K}(v_i(x_i) - c(x_i))a(x_i)\Big|\tilde{a}(x_i^*), i \in W_{x^*}\Big]$$

Note, that this is a special case of Definition 9 when assuming $L = \{S \subseteq x_i^*, i \in W_{x^*}\}$[1].

Knowing realized availabilities of allocated bundles, bidders have a total opportunity cost of $\sum_{i \in K}(v_i(x_i^*)\tilde{a}(x_i^*) - p_i)$ for joining coalition K. If they decide to join, then in expectation they can achieve the total welfare of $\omega_e(K)$ and thus they can provide at most $\omega_e(K) - \sum_{i \in K}(v_i(x_i^*)\tilde{a}(x_i^*) - p_i)$ of additional value to the seller, which gives rise to the following set of core constraints:

ECC-Core Constraint $\forall K \subset N$:

$$\sum_{i \in W_{x^*}}(p_i - c(x_i^*)\tilde{a}(x_i^*)) \geq \omega_e(K) - \sum_{i \in K}(v_i(x_i^*)\tilde{a}(x_i^*) - p_i)$$

By plugging these constraints as $Core^*$ into the mechanism framework together with $p^* = p^{\text{ECC-VCG}}$ as the reference point, we obtain a full specification of the *execution-contingent conditional core (ECC-Core)* mechanism. The following example demonstrates how a coalition imposes a core constraint:

Example 2. *Consider the setting from Example 1, with coalition $K = \{b_3, s\}$. The expected coalitional value for this coalition is $\omega_e(K) = 0.3 \cdot 0.4 = 0.12$. The corresponding core constraint is $p_1 + p_2 \geq 0.12$.*

4.5 ECR-Core Mechanism

If we know the realizations of the availabilities of *all* bundles at the consumption time, and not only of those *allocated*, then we can use more accurate execution-contingent core constraints:

ECR-Core Constraint $\forall K \subset N$:

$$\sum_{i \in W_{x^*}}(p_i - c(x_i^*)\tilde{a}(x_i^*)) \geq \omega(K) - \sum_{i \in K}(v_i(x_i^*)\tilde{a}(x_i^*) - p_i),$$

where $\omega(K) = \max_x \sum_{i \in K}(v_i(x_i) - c(x_i))\tilde{a}(x_i)$

Note that this definition is a special case of Definition 9 assuming $L = \{S : S \in 2^G\}$. To get a full specification of the mechanism we use these core constraints together with $p^{\text{ECR-VCG}}$ as parameters for the execution-contingent mechanism framework. As we will show in Section 5 exploiting this additional knowledge can significantly decrease the rate of IR violations.

The following theorem shows that ECC-VCG and ECR-VCG provide lower bounds for the corresponding core-selecting payment rules. This is useful to know, because it also implies that using these payment vectors as reference points in the overall mechanism framework makes sense.

Theorem 4. *ECC-Core and ECR-Core payments are lower-bounded by ECC-VCG and ECR-VCG payments respectively.*

Proof. Consider ECC-Core mechanism. If the ECC-Core is empty, then the mechanism charges ECC-VCG payments which are trivially lower-bounded by ECC-VCG payments. If the ECC-Core is not empty, then consider a coalition $K = N \backslash \{k\}$, where $k \in W_{x^*}$. In this case

$$\mathbb{E}[\omega(K)] = \mathbb{E}_f[SW(x^{-i})|\tilde{a}(x_i^*), i \in W_{x^*}]$$

Then,

$$p_i \geq \mathbb{E}_f[SW(x^{-i})|\tilde{a}(x_i^*), i \in W_{x^*}] -$$

$$\sum_{i \in W_{x^*} \backslash k}(v_i(x_i^*) - c(x_i^*))\tilde{a}(x_i^*) - c(x_k^*)\tilde{a}(x_k^*) = p_i^{\text{ECC-VCG}}.$$

The proof for ECR-Core is analogous. □

[1] Depending on the information structure of the domain, the availability that is revealed to the mechanism may only be $\tilde{a}(x_i^*), i \in W_{x^*}$ or $\tilde{a}(S), \forall S \subseteq x_i^*, i \in W_{x^*}$

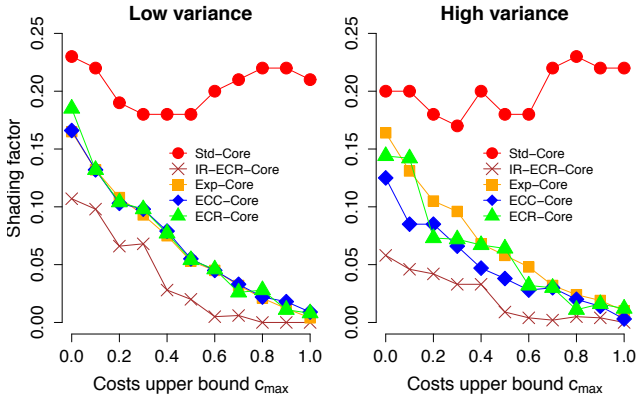

Figure 2: Additive shading factors of local bidders depending on costs distribution $c \sim U[0, c_{max}]$ for two levels of variance of the availabilities a.

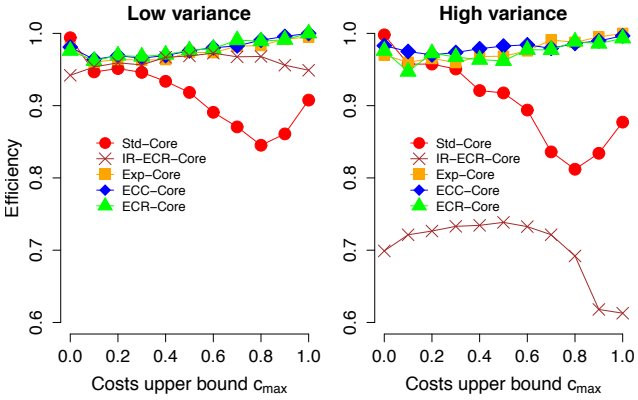

Figure 3: Efficiency depending on costs distribution $c \sim U[0, c_{max}]$ for two levels of variance of the availabilities a.

5 Comparison in Bayes-Nash Equilibrium

Our core-selecting payment rules are not strategyproof in expectation, and indeed, there do not even exist strategyproof core-selecting rules in domains without uncertainty [Goeree and Lien, 2016]. For this reason, we must analyze the properties (efficiency, IR violations, etc.) of our rules in equilibrium. The equilibrium concept we adopt for our auction domain is a *Bayes-Nash equilibrium*, i.e., where each bidder knows his own value, but only knows the distribution over other bidders' values, the seller's costs, and the availabilities.

Unfortunately, deriving the Bayes-Nash equilibrium of core-selecting payment rules analytically is very complex, and only feasible for very simple settings [Goeree and Lien, 2016]. For this reason we follow the approach by [Lubin and Parkes, 2009] and [Lubin *et al.*, 2016], and use a computational approach to find *approximate* BNEs for our rules. Concretely, we restrict the strategy space of the agents to *additive* shading strategies, and then use an algorithm based on fictitious play which, using an iterative best response method, converges to an ϵ-BNE in this restricted strategy space. Specifically, all the equilibria we report in this paper are ϵ-BNEs with ϵ=0.01.

5.1 Benchmark Rules

In addition to the ECC-Core and ECR-Core mechanisms, we also study the following three benchmark rules:

- **Std-Core Mechanism:** This refers to the standard core-selecting payment rule (the Quadratic rule as defined in [Day and Cramton, 2012]). The allocation rule selects an allocation assuming that all items are available, and payments are computed at allocation time.
- **Exp-Core Mechanism** This mechanism uses the ex-ante core as defined in Section 4.1, and then picks payments from this ex-ante core using the Quadratic rule [Day and Cramton, 2012]. Note that this mechanism satisfies individual rationality in expectation.
- **IR-ECR-Core Mechanism:** This refers to a rather artificial but still interesting mechanism. The allocation rule first maximizes expected social welfare, but then checks for every possible realization of the availabilities whether the resulting payments will satisfy ex-post IR. Only if this is the case will the allocation be made - otherwise no trade will happen. Thus, this rule is guaranteed to satisfy ex-post IR by design.

5.2 The Local-Local-Global (LLG) Domain

We study the well-known LLG domain. In this domain, there are two goods A and B. There are two local bidders who are each interested in one of the goods, and their values are drawn uniformly from $[0, 1]$, and there is a global bidders who is only interested in the bundle consisting of both goods, and his value for the whole bundle is drawn uniformly from $[0, 2]$. The seller's costs are drawn uniformly from $[0, c_{max}]$, where c_{max} is a parameter which we vary in our analysis.

It is easy to show that under Exp-Core, ECC-Core, ECR-Core, and IR-ECR-Core, if the global bidder gets allocated, he is charged the respective generalized version of the VCG payment. Thus, by Theorem 1, ECC-Core and ECR-Core are strategyproof in expectation for the global bidder, and the same result can be shown for Exp-Core and IR-ECR-Core. For this reason, we only need to compute the BNE strategies of the local bidders for those rules. For Std-Core, we also compute the BNE strategy for the global bidder.

5.3 Results

Strategies. Figure 2 shows the BNE strategies of the local bidders; on the left side for a domain with low variance in the availabilities of the goods, and on the right side for high variance in the availabilities of the goods. Note that these results are not "simulations," but that each point in these figures is the result of our BNE algorithm (and it takes about 8 hours to find a BNE for one payment rule on a machine with 20 cores).

We see that Std-Core has the worst incentives; the global bidder actually also shades (not shown in Figure 2), with a shading factor roughly twice as high as that of the local bidders. ECC-Core, ECR-Core and Exp-Core have very similar incentives, and IR-ECR-Core has the best incentives. Furthermore, the higher the costs, the lower the shading factors (except for Std-Core), which makes sense, because with higher costs, the opportunities for trade get smaller, and thus it gets more risky

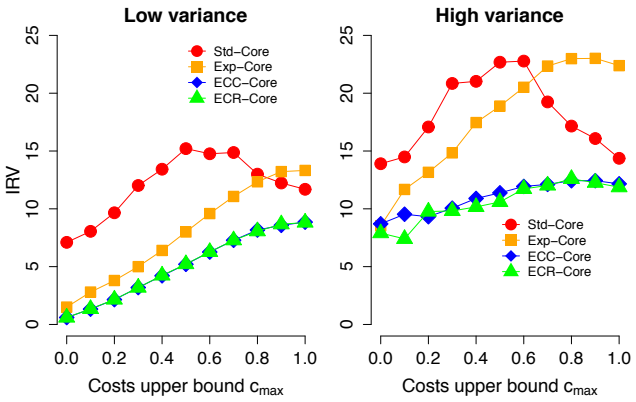

Figure 4: Rate of IR violations (%) depending on costs distribution $c \sim U[0, c_{max}]$ for two levels of variance of the availabilities a.

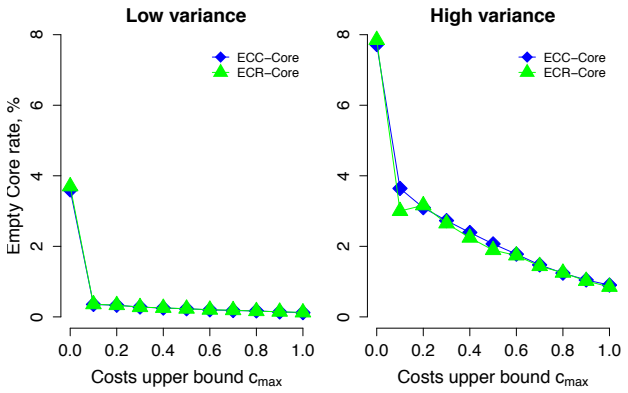

Figure 5: Rate empty core cases (%) depending on costs distribution $c \sim U[0, c_{max}]$ for two levels of variance of the availabilities a.

to shade. We also see that the shading factors are lower in the high variance domain, which makes sense, because the higher the uncertainty, the more risky it is to shade.

Efficiency. Figure 3 shows the efficiency achieved by all mechanisms (we simulate 1 million auctions using the computed BNE strategies to calulate the efficiency). We see that our new mechanisms achieve higher efficiency than both, Std-Core as well as IR-ECR-Core.[2] By comparing the low variance and high variance domains, we see in what sense the IR-ECR-Core mechanism is really just a straw-man: while it works well in the low variance domain, in the high variance domain the mechanism has to cancel a lot of allocations because it cannot guarantee ex-post IR for all possible realizations of the availabilities, which drives down efficiency. In contrast, our execution-contingent mechanisms can handle this uncertainty.

Rate of Ex-post IR Violations. Figure 4 shows the rate of ex-post IR violations. Again, Std-Core performs worst. But this analysis now also demonstrates the advantages of ECC-Core and ECR-Core over Exp-Core. While all three mechanisms had good incentives and high efficiency, we now see that ECC-Core and ECR-Core have a significantly lower rate of ex-post IR violations, which makes sense because ECC-Core and ECR-Core are execution-contingent. This advantage is even more pronounced for the high-variance domain. In this domain, we even see a small advantage of ECR-Core over ECC-Core, which makes sense, because ECR-Core takes even more information into account when computing payments.

Empty Core Analysis. While in a domain without uncertainty, there always exists a price vector in the core, one of the interesting and perhaps unexpected features of our domain is that the ECC-Core as well as the ECR-Core can be empty. For this reason, we also study how often this happens in our domain. Figure 5 shows that in the low variance domain, the core is typically not empty, especially if the costs are non-zero then the probability of an empty core is close to 0. However, in the high variance domain, the rate of empty core cases is significantly higher. This is explained by the fact that the higher the variance in availabilities of bundles, the more flexible the core constraints are, and thus the more likely they can be in conflict with each other (and thus lead to an empty core).

[2]There is one exception: with zero costs, Std-Core achieve 99% efficiency while our rules achieve 98% efficiency. This happens because under our rules, the global bidder plays truthful and the local bidders shave, which leads to an efficiency loss. Under Std-Core, all bidders shave roughly proportionally, which leads to almost no efficiency loss in equilibrium (with zero costs). However, this peculiarity of the LLG domain does not generalize to larger domains.

6 Conclusion

In this paper, we have studied mechanisms for combinatorial auctions with uncertain availabilities of goods. We have introduced two execution-contingent core-selecting payment rules which both satisfy IR in expectation. Furthermore, we have performed an extensive computational Bayes-Nash equilibrium analysis, comparing our new rules with three benchmark rules to study the trade-off between different mechanism design objectives. Our results show that, compared to a standard core-selecting auction, our rules have significantly higher efficiency and lower ex-post IR violations. Furthermore, comparing our two execution-contingent mechanisms, we conclude that the more information about realized availabilities the mechanism has, the more of it should be exploited in the computation of the payments, as this leads to an lower rate of IR violations. Moreover, by comparing our rules to an ex-post IR rule, we have shown that by relaxing the strict ex-post IR constraint, we can gain a lot in efficiency. Thus, if a small rate of ex-post IR violations is acceptable, then we recommend using one of our new payment rules for a combinatorial auction domain with uncertain availabilities of goods.

Acknowledgments. We would like to thank Timo Mennle for insightful discussions and the anonymous reviewers for helpful comments. This research is supported by the SNSF (Swiss National Science Foundation) project #153598.

References

[Ausubel and Baranov, 2014] Lawrence Ausubel and Oleg Baranov. A practical guide to the combinatorial clock auction. Working paper, 2014.

[Ausubel and Baranov, 2016] Lawrence M. Ausubel and Oleg V. Baranov. Core-selecting auctions with incomplete information, 2016. Working paper.

[Ausubel and Milgrom, 2006] Lawrence M. Ausubel and Paul Milgrom. The lovely but lonely vickrey auction. In Peter Cramton, Yoav Shoham, and Richard Steinberg, editors, *Combinatorial Auctions*, chapter 1, pages 57–95. MIT Press, 2006.

[Ceppi *et al.*, 2015] Sofia Ceppi, Nicola Gatti, and Enrico H. Gerding. Mechanism Design for the Federation of Service Providers. Working Paper, 2015.

[Clarke, 1971] E. H. Clarke. Multipart Pricing of Public Goods. *Public Choice*, 2:19–33, 1971.

[Cramton and Doyle, 2015] Peter Cramton and Linda Doyle. An open access wireless market supporting public safety, universal service, and competition. Technical report, 2015.

[Cramton, 2013] Peter Cramton. Spectrum auction design. *Review of Industrial Organization*, 42(2):030–190, March 2013.

[Day and Cramton, 2012] Robert Day and Peter Cramton. Quadratic core-selecting payment rules for combinatorial auctions. *Operations Research*, 60(3):588–603, 2012.

[Day and Milgrom, 2008] Robert Day and Paul Milgrom. Core-selecting Package Auctions. *International Journal of Game Theory*, 36(3):393–407, 2008.

[Day and Raghavan, 2007] Robert W. Day and S. Raghavan. Fair payments for efficient allocations in public sector combinatorial auctions. *Management Science*, 53(9):1389–1406, September 2007.

[FCC, 2016] https://www.fcc.gov/general/700-mhz-public-safety-spectrum-0#block-menu-block-4; last accessed on: February 2, 2016.

[Goeree and Lien, 2016] Jacob Goeree and Yuanchuan Lien. On the impossibility of core-selecting auctions. *Theoretical Economics*, 2016. Forthcoming.

[Goetzendorf *et al.*, 2015] Andor Goetzendorf, Martin Bichler, Pasha Shabalin, and Robert W. Day. Compact Bid Languages and Core Pricing in Large Multi-item Auctions. *Management Science*, 61(7):1684–1703, 2015.

[Groves, 1973] T. Groves. Incentives in Teams. *Econometrica*, 41(4):617–631, 1973.

[Lubin and Parkes, 2009] Benjamin Lubin and David C. Parkes. Quantifying the strategyproofness of mechanisms via metrics on payoff distributions. In *Proceedings of the 17th National Conference on Artificial Intelligence*, pages 74–81, 2009.

[Lubin *et al.*, 2016] Benjamin Lubin, Benedikt Bünz, and Sven Seuken. New Core-Selecting Payment Rules with Better Fairness and Incentive Properties. Working Paper, 2016.

[Milgrom, 2007] Paul Milgrom. Package auctions and exchanges. *Econometrica*, 75(4):935–965, 2007.

[Moor *et al.*, 2015] Dmitry Moor, Tobias Grubenmann, Sven Seuken, and Abraham Bernstein. A Double Auction for Querying the Web of Data. In *Proceedings of the Third Conference on Auctions, Market Mechanisms and Their Applications (AMMA)*, Chicago, USA, August 2015.

[Porter *et al.*, 2008] Ryan Porter, Amir Ronen, Yoav Shoham, and Moshe Tennenholtz. Fault tolerant mechanism design. *Artificial Intelligence*, 172(15):1783 – 1799, 2008.

[Ramchurn *et al.*, 2009] Sarvapali D. Ramchurn, Claudio Mezzetti, Andrea Giovannucci, Juan A. Rodriguez-Aguilar, Rajdeep K. Dash, and Nicholas R. Jennings. Trust-based mechanisms for robust and efficient task allocation in the presence of execution uncertainty. *Journal of Artificial Intelligence Research*, 35(1):119–159, June 2009.

[Sandholm, 2013] T. Sandholm. Very-large-scale generalized combinatorial multi-attribute auctions: Lessons from conducting $60 billion of sourcing. In Alvin E. Roth Nir Vulkan and Zvika Neeman, editors, *The Handbook of Market Design*, chapter 1. Oxford University Press, 2013.

[Seuken *et al.*, 2015] Sven Seuken, Pertti Juhani Visuri, Johanna Katariina Visuri, and Dan Zagursky. Systems and methods for allocating and pricing alternative network access resources with reserve prices, July 2015, U.S. Patent 0189580.

[Vickrey, 1961] William Vickrey. Counterspeculation, Auctions, and Competitive Sealed Tenders. *The Journal of Finance*, 16(1):8–37, 1961.

4 Designing a Marketplace for Distributed Data

> "But all profit that is out of proportion to the labor expended is dishonest."
> "But who is to define what is proportionate?"
>
> Leo Tolstoy, Anna Karenina

The content of this chapter has previously appeared in:

> Moor, D. and Seuken, S. and Grubenmann, T. and Bernstein, A. (2019). Designing a Marketplace for Distributed Data. *working paper*.

The Design of a Combinatorial Data Market

Dmitry Moor dmoor@ifi.uzh.ch
Sven Seuken seuken@ifi.uzh.ch
Tobias Grubenmann grubenmann@ifi.uzh.ch
Abraham Bernstein bernstein@ifi.uzh.ch
University of Zurich, Binzmuhlestrasse 14,
Zurich, 8050 Switzerland

Abstract

In this paper, we propose a model of a production economy of data markets. Such markets typically exhibit a complex production structure when the databases produced by different data providers can be joined to produce answers for buyers' queries. We focus on three specific challenges: (1) different providers have the capability to produce different sets of databases; (2) data providers have high fixed costs for producing a database; and (3) buyers have combinatorial values over which databases are produced and thereby become available in the marketplace. To illustrate our model, we provide a possible market design solution for this domain. The key idea of our solution is to use a reverse auction for the sellers, a posted-price mechanism for the buyers, and a fixed-point iteration algorithm for finding an outcome that balances the two sides of the market. To achieve this, we illustrate how to infer values of buyers for particular databases from the values for their queries. Via simulations, we show how our market distributes the surplus between buyers and sellers. In particular, we demonstrate that our design rewards providers of "unique" data much more than providers of "common data."

1. Introduction

Many datasets on the Web are unstructured, i.e., they can be interpreted by humans but not by machines. There are numerous domains in which we would benefit greatly from data published in a *structured* way, for example, as a database. This allows machines to understand relationships between different pieces of data (Bernstein et al., 2016). Consequently, this significantly reduces the effort for humans to analyze lots of unstructured datasets that are discovered by traditional search engines. Instead, one could delegate this task to automatic query processing algorithms. For example, in the life sciences, researchers submit queries that join data from multiple databases provided by different companies. Each of these databases contains information on chemical compounds, disease data, biological function, and biomarkers. Automatic aggregation and processing of this data leads to faster and more efficient drug discovery (HCLS, 2001). Another example is IBM Watson, a large scale question answering system that defeated the human champions in the well-known TV show *Jeopardy*. This system heavily relies on querying structured data from distributed databases (Ferrucci et al., 2010) and has numerous applications in cancer treatment and clinical research, financial advisory, and retail.

Technology that enables automatic aggregation and processing of structured data already exists, for example, the *Web of Data* (WoD) (W3C, 2014). This technology does all the

work of joining and processing the data and returns a precise answer to the user's query. However, despite the apparent power of the WoD approach, the technology has not yet seen widespread adoption. One of the reasons for this underutilization is of economic nature: most of the data produced for the WoD so far was either subsidized by governments or produced at a loss (Buil-Aranda et al., 2013). This suggests that one of the most important reasons preventing wide adoption of the WoD is a lack of financial incentives for data providers to publish their data in a structured way. Indeed, data providers may incur high costs for producing their databases, i.e., for structuring their data and linking it against databases of other data providers. Naturally, data providers hope to recoup these costs. However, advertisement, the main source of income for many data publishers as well as traditional search engines, does not work in the Web of Data because in the WoD, data is processed by machines rather than by humans and the machine can simply ignore any ad. Therefore, new sources of revenue for data providers are needed. One possible way to achieve this is via a market in which providers sell data to users and trade is mediated by a market platform. In this paper, we propose a model of the economy that enables to implement such a market.

1.1 Call for Data Markets

The need for data markets was recognized by both business and academic communities. In a recent McKinsey report (2016), for example, the authors explained the need for data markets by referring to the inefficient use of constantly increasing amounts of data produced by businesses adopting IoT technologies.

Schomm et al. (2013) provide a good overview of existing data markets. One prominent practical data market was operated by Microsoft with the Microsoft Azure Data Marketplace platform, but seized operation in 2016 due to a "lack of sustained customer interest" (Ramel, 2016). This lack of interest, however, does not imply a lack of demand for data. A more likely explanation is an inadequate business model. This explanation is supported by the fact that companies like Thomson Reuters, LexisNexis and Bloomberg still make large profits by selling access to their proprietary databases (Thomson-Reuters, 2015; Greg, 2011). However, it is not possible to easily combine and process their data with data from databases produced by other data providers.

There are numerous challenges when designing markets for *information goods* such as data. Already more than 20 years ago, Varian (1995, 1997) highlighted the problem of high sunk and low marginal production costs for these goods. Bakos and Brynjolfsson (1999) studied the problem that buyers may have high uncertainty regarding their valuations for information goods. More recently, Moor et al. (2015) and Grubenmann et al. (2018a) argued that, in data markets, the combinatorial preferences of buyers should be taken into account when the buyers are able to join multiple databases. This idea of combinatorial preferences is further emphasized by Agarwal et al. (2019) who suggested a possible solution for a marketplace that sells data for machine learning tasks. In their paper, the authors proposed a design of a market for a setting when data providers have already incurred high sunk costs for producing their data and would naturally like to minimize their regret. In contrast to that, in our work, we aim at designing such a market that would incentivize sellers who

have not yet produced their data to do so. In particularly, we want to guarantee that these sellers always get fully compensated for producing their databases. [1]

1.2 Overview of our Approach

In this paper, we propose a game theoretical model of the production economy of data markets. We focus on three challenges: (1) databases are distributed (produced by different data providers) and can be joined to produce an answer to a query; (2) data providers have high fixed costs for producing a database; and (3) buyers have combinatorial values over which databases are available (i.e., different combinations of databases lead to more or less valuable answers to a buyer's query).

Goldberg et al. (2001), Goldberg and Hartline (2001, 2003) also studied markets for information goods. They proposed an auction for selling goods in unlimited supply (such as data) in a setting with a single seller. However, they did not consider a setting with multiple distributed sellers or buyers with combinatorial values.

In recent papers by Balazinska et al. (2013), Koutris et al. (2013, 2015) and Deep and Koutris (2016), the authors aimed at designing a data market with quoted prices. The basic idea was to charge a different price for different *views* of the database in a way that would guarantee a *no-arbitrage* property. However, their approach does not allow joining data from multiple different data providers and ignores the costs of production.

While many authors (e.g., Varian (1995), Goldberg et al. (2001)) have previously studied the economic problem of how to sell an information good (such as music files or videos), their approaches do not translate to data markets. The main reason is the combinatorial structure that arises in a data market once we allow databases to be joined: a buyer's value for receiving answers based on one database may be zero while it may be very large once two databases are joined. This combinatorial structure is not present with music or video files which is why data markets require a new design.

The general approach we adopt to model such a two-sided market is as follows. First, we design an auction to elicit the data providers' production costs. The objective of the auction is to allocate data providers (and respective databases) in a way that maximizes the total utility of buyers. This needs to be done subject to the constraint that the production costs of data providers are recouped. Second, we suggest a uniform posted pricing scheme that is presented to buyers submitting their queries. Apart from its simplicity, the use of uniform posted prices prevents complex strategic behavior of buyers who are assumed to be price-takers. Finally, we propose a mechanism that makes the overall market budget balanced, i.e., that guarantees that the total expected amount of money collected from buyers is equal to the total payment that needs to be accrued to data providers. While there are many possible equilibria that can arise in such a market (for example, a trivial equilibrium where nobody is allocated), our mechanism targets at finding the one with the highest surplus. This balancing mechanism ties together with the reverse auction on the one side and the posted price mechanism on the other side.

1. These two different market design objectives result in different amounts of information that the market designer needs to know to compute allocation and payments. In particular, the design of Agarwal et al. (2019) does not rely on prior knowledge of distributions of sellers' costs while our design does.

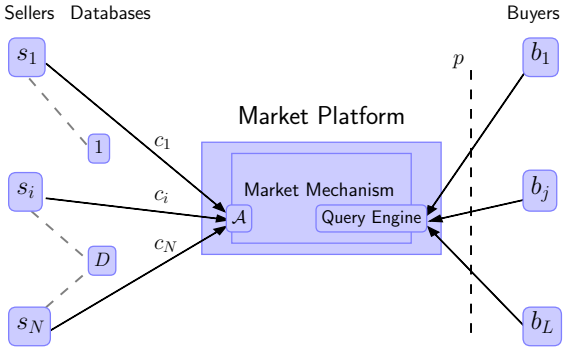

Figure 1: Market structure. Buyers consume rows of database tables corresponding to their queries for a posted price p per row. The reverse auction $\mathcal{A}$ elicits fixed costs of sellers. The market platform balances both sides of the market, i.e., the total payment collected from buyers needs to be equal to the total payment accrued to sellers.

Scope of this work. To keep our model simple, we focus only on the most essential economic features of the domain such as high fixed costs and low marginal costs of production of data, the distributed nature of production and the combinatorial aspect of users' preferences. We argue that these features have the most significant implications on market design. The list of all possible features may include *privacy* and *quality* of data, dynamic arrival of sellers to the market etc. Accounting for privacy and quality of data can make the model unreasonably complicated and are therefore outside the scope of the current paper. Dynamic arrival of sellers may play an important role in designing combinatorial markets. This follows from the fact that an allocation of a seller today depends on whether another seller with a complementary database arrives tomorrow. However, even our *static* model already enables us to capture most of the interesting aspects of practical data markets while remaining relatively simple. Thus, we leave the dynamic arrival of sellers to be studied in future extensions of this work.

2. Preliminaries

2.1 Formal Model

Figure 1 provides a schematic overview of the market we will design. We assume that there are N sellers (i.e., data providers), $s_1, ..., s_N$; and L buyers (or users), $b_1, ..., b_L$, who submit their queries; $N, L > 0$ are given exogenously. To keep the model simple, we assume that each buyer submits a single query. The answer to a buyer's query consists of multiple rows of a database table (possibly joined over multiple databases) that satisfy the buyer's query. Thus, the buyer can buy *some* of these rows in exchange for money. This means that there are two goods involved in the exchange: *money* and *rows* of database tables corresponding to buyers' queries.

Market Structure. The market is mediated by a *market platform*. This market platform operates a so-called *query engine*, which executes buyers' queries and produces answers for

these queries. Given a query from buyer b_j, the query engine takes databases of different sellers as inputs and then outputs $R_j \in \mathbb{N}$ rows by joining those databases.

In our market design, the market platform is a *neutral* (and, in particular, non-profit maximizing) entity for two primary reasons. First, observe that there are two levels of production: the sellers produce their databases, and the query engine then takes/joins those databases to produce answers to buyers' queries. Thus, it is highly convenient to consider the query engine a separate entity residing at the market platform. Second, one can show that in a domain with zero marginal costs of production (such as our domain), a (non-trivial) competitive equilibrium (i.e., where the seller maximizes her profits and the buyer maximizes his utility) is not guaranteed to exist (Mas-Colell et al., 1995).[2] Similar problems occur in markets with natural monopolies (Tirole & Laffont, 1993). In practice, such markets require a *regulator*, and the regulator is usually a governmental organization that induces prices based on its own analysis of production costs and market demand. This fact comprises our second argument for the use of the neutral (non-strategic) market platform that acts as a ''regulator.''

In our case, it is the market platform that ''sets prices" based on two key factors: the sellers' production costs and the market demand for the databases. To elicit the production costs of the sellers we argue for the use of a reverse auction $\mathcal{A}$ (defined formally later in this section). We also argue for the use of a uniform posted price *per row* that is exposed to buyers for estimating the demand for databases (see Figure 1).

Sellers. We assume that every seller can produce a single database and that $c_i \in \mathbb{R}_{\geq 0}$ is the fixed cost of s_i for producing her database.[3][4] For the sake of simplicity we assume that different databases are disjoint. Let $c = (c_1, ..., c_N)$ be the cost profile of sellers. Let D be the number of different databases, $D \leq N$. For every database $k \in \{1, ..., D\}$ we let $i(k) = (i_1(k), ..., i_q(k))$ denote the indices of sellers that can produce the database k.

We assume that the c_i are independent random variables distributed according to cumulative probability distributions F_i, $i = 1, ..., N$. We let f_i be the corresponding probability density. We assume that $f_i(c_i)$ has full support on some interval $[\alpha_i, \beta_i]$. Then, the joint probability density is $f(c) = \prod_{i=1}^{N} f_i(c_i)$. Similarly, $f_{-i}(c_{-i}) = \prod_{j \neq i} f_j(c_j)$ is the joint probability density of all sellers except s_i.

Let $t = (t_1, ..., t_N)$ denote transfers (payments) received by s_i and $a = (a_1, ..., a_N)$ denote an allocation decision of the reverse auction $\mathcal{A}$, i.e., $a_i \in [0, 1]$ is the probability that s_i is allocated. [5] The utility of seller s_i is assumed to be quasi-linear, i.e, $u_i(a, c_i, t) = t_i - a_i c_i$.

Sellers are strategic and can thus misreport their costs. Let $\hat{c}_i$ denote the reported cost of s_i. Then, $(\hat{c}_1, ..., \hat{c}_N)$ is a reported cost profile of all sellers. Similarly, $\hat{c}_{-i} = (\hat{c}_1, ..., \hat{c}_{i-1}, \hat{c}_{i+1}, ..., \hat{c}_N)$ denotes a reported cost profile of all sellers except s_i.

Buyers. We assume that every buyer b_j is equipped with an initial *endowment* of money $e \in \mathbb{R}_{\geq 0}$. A buyer can use his endowment to acquire rows of the database table corresponding

2. This follows from the fact that the profit maximizing seller will produce the maximum possible number of rows at zero marginal cost if the price per row is positive. The buyer, however, may not be willing to pay for all these rows unless the price is 0.
3. We assume marginal costs, i.e., costs of maintaining a database and answering queries, to be zero.
4. Throughout the paper we use ''she'' for sellers and ''he'' for buyers.
5. For technical reasons and simplicity of some proofs we assume that the allocation is probabilistic. The resulting mechanism that we will present in Section 3 however, will be deterministic.

to his query and keep the rest of this money, $m_j \in \mathbb{R}_{\geq 0}$, in his wallet. Let $r_j \in \mathbb{N}$ denote the number of rows acquired by b_j.

We assume that the buyers in our market are *institutional* agents, i.e., agents which are buying data for an institution and not an individual person. They could be pharmaceutical companies, operators of cloud applications or some other kind of intermediary. Importantly, we assume that the buyers in our domain can estimate their value for rows of the database tables. We assume that buyers are risk-neutral the b_j's preferences are described by a quasi-linear utility function $u_j(m_j, r_j, a) = v_j(r_j, a) + m_j$. Here, v_j is the value function of b_j. Notice that the value function depends on the allocation decision a of $\mathcal{A}$ regarding data providers. This follows from the assumption that the larger the number of allocated sellers, the more "informative" (and thus valuable) an answer for the buyer's query becomes. For simplicity, we assume that for any a, $v_j(r_j, a)$ is linear and non-decreasing in r_j up to a certain threshold $\overline{r}_j(a) \geq 0$, and exhibits zero marginal increase for every additional row $r_j > \overline{r}_j(a)$ (see Figure 2).[6] Depending on the allocation a, each buyer can decided to buy less or more data, and thus, the threshold $\overline{r}_j(a)$ depends on the allocation a. Shortly, we will elaborate on this dependency. Formally, $v_j(r_j, a) = v'_j(a) \min\{r_j, \overline{r}_j(a)\}$. Here, $v'_j(a)$ is the marginal value for rows if the allocation is a. We let $F_{v'(a)}$ and $F_{\overline{r}(a)}$ denote the cumulative distribution functions of $v'_j(a)$ and $\overline{r}_j(a)$, respectively.

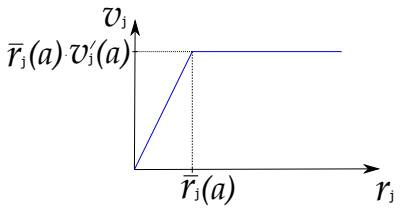

Figure 2: Value function of the buyer b_j. Here, $\overline{r}_j(a)$ and $v'_j(a)$ are the threshold and the marginal value of b_j when the allocation is a.

We make two assumptions regarding how buyers' preferences change when additional sellers (and thus, additional databases) are allocated. Assume that $a = (a_1, ..., a_i, ..., a_N)$ is an allocation and let us define $\delta a_i = (0, ..., \delta, ..., 0)$ with the ith element equal to δ, $0 \leq \delta \leq 1 - a_i$.

Assumption 1 (**Monotonicity of the Marginal Value**). *For every $j = 1, ..., L$, $\forall i = 1, ..., N$ the following inequality holds $v'_j(a + \delta a_i) \geq v'_j(a)$.*

The intuition behind this assumption is as follows: As more sellers are allocated, the marginal value of every buyer b_j for the answer of his query cannot decrease. This assumption is justified by the fact that the more information there is available, the more precise the answer to the query will be. Indeed, joining two or more databases implies computation of a certain relational predicate on the data. This results in obtaining conditional distributions over the data and consequently, in more precise answers.[7]

Assumption 2 (**Monotonicity of the Maximum Buyer's Value**). *For every $j = 1, ..., L$, $\forall i = 1, ..., N$ it holds $v'_j(a + \delta a_i)\overline{r}_j(a + \delta a_i) \geq v'_j(a)\overline{r}_j(a)$.*

6. Generally speaking, buyers could have a decreasing marginal value for additional rows. However, the assumption of a constant marginal value and a threshold is not too restrictive as it still captures convex preferences of buyers while providing us with a relatively simple model.
7. While this model captures a number of different JOIN operations over the databases (e.g., inner JOIN, outer JOIN, etc.), it may seem less applicable for UNION operations. Our model resolves by treating a query that takes the union of two databases as two separate queries.

This assumption can be interpreted as follows. Buyers submit their queries in order to answer a particular question. As answers for queries become more valuable (due to Assumption 1), each buyer can decide to buy more or fewer rows. However, this doesn't change the value of answering the particular question they had in mind: more information simply permits a better approximation of the question because it makes use of more databases. This makes the total value of a query answer, $v'_j(a)\overline{r}_j(a)$, non-decreasing.

Finally, we assume that buyers are indifferent between identities of sellers who produce a database k.

Allocation and Pricing. Let p denote the posted price per row in an answer to a query. The price p will be set independent of the query to prevent buyers from engaging in complex strategic behavior when deciding which queries to submit (e.g., to get cheaper answers to their questions). While this may seem counterintuitive at first sight, remember that producing any answer has zero marginal costs for the seller. Further, we assume the number of buyers L to be large, such that buyers are price takers and thus cannot manipulate the market price.

Given price p and allocation probabilities a, every buyer b_j solves his *consumption problem* of maximizing his utility subject to the budget constraint (Mas-Colell et al., 1995):

$$\begin{aligned} &\max_{m_j, r_j} u_j(m_j, r_j, a) \\ &\text{s.t. } p \cdot r_j + m_j \leq e, \\ &\quad m_j \geq 0, r_j \leq R_j. \end{aligned} \tag{1}$$

Let $(m_j^*(p,a), r_j^*(p,a))$ be a solution to the consumption problem when the posted price is p and the allocation is a. Here, $r_j^*(p,a)$ is the *(Marshalian) demand of* b_j for rows when the posted price is p and the allocation is a. Similarly, $m_j^*(p,a)$ is the demand of the buyer for money (i.e., how much money the buyer wants to keep). Observe that $m_j^*(p,a)$ and $r_j^*(p,a)$ need not be functions. In fact, they are correspondences (Mas-Colell et al., 1995).

Let $\mathcal{A} = \langle g, h \rangle$ be the reverse auction adopted by the market platform. Here, $g : \mathbb{R}^N \to [0,1]^N$ denotes an allocation rule that maps the cost profile $c = (c_1, ..., c_N)$ to the allocation decision $(a_1, ..., a_N)$; $a_i = g_i(c_i, c_{-i})$, where $g_i(c_i, c_{-i}) : \mathbb{R}^N \to [0,1]$ computes the probability that s_i is allocated. If a is an allocation, we let $\lfloor a \rfloor_k$ be the corresponding allocation in which sellers producing database k are not allocated. We let h denote the payment rule that maps the cost profile $c = (c_1, ..., c_N)$ to the vector of payments $(t_1, ..., t_N)$ to be paid to the sellers. As we shall see in Section 3, these payments need to be made "in expectation" due to the random nature of buyers' values and sellers' costs. As we discuss in Section 3, when $\mathcal{A}$ computes the allocation and payments it takes into account F_i for all $i = 1, ..., N$ as well as both $F_{v'(a)}$ and $F_{\overline{r}(a)}$ for all possible deterministic allocations $a \in \{0,1\}^N$ of sellers.[8]

Given all of the above information, the market platform can compute the price per row $p \geq 0$, which is exposed to buyers, and payments $t_i \geq 0$ to be paid to sellers.

8. Observe that as the number of databases increases, the number of possible deterministic allocations increases exponentially. In practice, the valuations of buyers for different allocations may be very similar. This would allow the market platform to considerably reduce the amount of information it needs to collect and this would also simplify pricing. In this paper however, we do not discuss such optimizations.

Remark 2.1. *In practice, the market platform will need to do some market research to learn the distributions $F_i, F_{v'(a)}, F_{\overline{r}(a)}$. Such learning can be done by building an appropriate regression model that captures the connectivity of different databases, their topics, the validity of the data, etc. The design of an appropriate learning procedure, however, is outside of the scope of this paper.*

Finally, we define the social welfare as the total utility of buyers and sellers obtained in the market given allocation a, payments t and the posted price p, i.e., $SW(a,t,p) = \sum_{i=1}^{N} u_i(a,c_i,t) + \sum_{j=1}^{L} u_j(m_j^*(p,a), r_j^*(p,a), a)$.

2.2 Market Properties

We now discuss a number of properties we would like the reverse auction $\mathcal{A}$ and the overall market mechanism to satisfy.

Auction Properties. We begin with the properties we would like the reverse auction $\mathcal{A}$ to satisfy.

Definition 1. *The reverse auction $\mathcal{A} = \langle g, h \rangle$ is **Bayes-Nash incentive compatible** (BNIC), if $\forall i = 1, ..., N$ $\forall c_i$ $\forall \hat{c}_i$ $\forall c_{-i}$*

$$\mathbb{E}_{f_{-i}}[u_i(g(c_i, c_{-i}), c_i, h(c_i, c_{-i}))] \geq \mathbb{E}_{f_{-i}}[u_i(g(\hat{c}_i, c_{-i}), c_i, h(\hat{c}_i, c_{-i}))]. \tag{2}$$

In our work, we look for a reverse auction $\mathcal{A}$ that satisfies BNIC.

The following property guarantees participation of sellers in the reverse auction:

Definition 2. *The reverse auction $\mathcal{A} = \langle g, h \rangle$ is **individually rational** (IR) for sellers, if $\forall i = 1, ..., N$, $\forall c_i, \forall c_{-i}$*

$$\mathbb{E}_{f_{-i}}[u_i(g(c_i, c_{-i}), c_i, h(c_i, c_{-i}))] \geq 0. \tag{3}$$

Market Mechanism Properties. Now, we switch to a discussion of the properties that the overall market mechanism should have.

First, observe that individual rationality is satisfied for buyers automatically. This follows from the fact that when solving their consumption problem (1), buyers always have the option not to consume rows and to keep their whole endowment e.

Additionally, we would like the market mechanism to be budget balanced. Formally:

Definition 3. *The market mechanism is **budget balanced** (BB) if $\forall c$, $\forall F_{v'(a)}, F_{\overline{r}(a)}$, $\forall F_i$ $(i = 1, ..., N)$, the price p, the allocation and payments computed by $\mathcal{A} = \langle g, h \rangle$ satisfy*

$$\sum_{i=1}^{N} t_i = \sum_{j=1}^{L} (e - m_j^*(p, a)), \tag{4}$$

where $(t_1, ..., t_N) = h(c)$ and $a = g(c)$.

In words, the total payment to sellers computed by $\mathcal{A}$ should be equal to the total amount of money collected from buyers. As $\mathcal{A}$ uses $F_{v'(a)}, F_{\overline{r}(a)}$ and F_i, $i = 1, ..., N$ to compute the allocation and payments (see Section 2.1), this property should hold for arbitrary distributions.

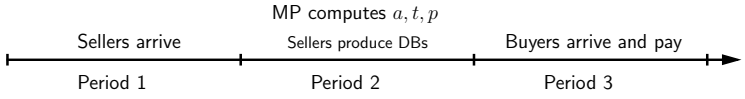

Figure 3: Temporal model of the market. During the first time period, sellers arrive and report their costs. In the second time period, the market platform computes an allocation and payments to sellers, as well as the posted price p. During the third time period, buyers arrive and consume the desired number of rows of database tables corresponding to their queries.

2.3 Temporal Structure

Figure 3 illustrates the temporal structure of the data market we propose. First, for simplicity, we present the model with only three time periods and then elaborate on how it can be generalized for an arbitrary time horizon T.

In time period $\tau = 1$, all sellers s_i, $i = 1, ..., N$ arrive to the market and report their costs to the market platform. Then, in time period $\tau = 2$, the market platform computes allocation $a = (a_1, ..., a_N)$ and payments $t = (t_1, ..., t_N)$ as well as the posted price per row, p, based on reported costs received from sellers and the value model of buyers (i.e., $F_{v'(a)}$ and $F_{\overline{r}(a)}$). Once the allocation is computed, allocated sellers can produce their databases (this happens during the same time period $\tau = 2$). Finally, in time period $\tau = 3$, buyers arrive to the market, submit their queries and pay a price p per row of answers to their queries.

This simple model can be generalized straightforwardly to a setting where buyers do not arrive to the market at the same time, but over a certain time horizon T. In this setting, we assume that each database remains fully relevant (i.e., the value of the buyers remains the same) for T time periods, but that the data in the database becomes obsolete and thus has zero value after T time periods.[9] This means that we assume that T is the *timeliness* of data (i.e., a period of time during which the data is still up to date).

The market platform promises to the sellers that over the time horizon T, every allocated seller s_i will receive payments that are expected to add up to t_i.[10] This means that the market platform does not pay t_i to the seller s_i immediately at time $\tau = 1$. Instead, it will accrue money paid by buyers coming to the market over the time horizon T. We assume that L buyers arrive to the market over the time horizon T, each paying $e - m_j^*(p, a)$ to the market platform to get $r_j^*(p, a)$ rows of database tables (possibly joined) corresponding to their queries. These payments go directly to sellers until all promises are fulfilled (i.e., each seller s_i receives t_i).

9. It is also possible to use a discounting factor for the value of data. We leave this direction for future work.
10. In our model, we assume that exactly L buyers arrive. In practice, more or fewer buyers would arrive. Our model extends straightforwardly to these cases by including an additional expectation over the number of buyers.

3. Market Design

In this section, we present our main contribution: a market design solution for selling distributed data. We demonstrate how the reverse auction $\mathcal{A}$ should be designed as well as how the posted price p must be computed to achieve our design goals.

3.1 Market Design Objective and Constraints

Social Choice Function. The result of Myerson and Satterthwaite (1983) in domains when buyers and sellers are both strategic implies that there does not exist a social welfare maximizing mechanism that is BNIC, IR and budget balanced. Thus, optimizing social welfare is not feasible in this domain. Instead, we find it acceptable to sacrifice a bit of social welfare as long as the resulting market is BNIC and guarantees IR and BB. Given this, we can consider either optimizing the total utility of buyers or the revenues of sellers, subject to the aforementioned constraints.

As discussed in Section 1.1, non-negative profits for sellers is a crucial constraint for the viability of markets for distributed data. This is why maximizing the revenue of sellers could be a potential objective. This objective, however, does not seem very attractive, as we envision the market for distributed data to give rise to many novel AI applications coming from the buyers' side. Distributing all the surplus in favor of the sellers (who are often "monopolists" of their data) can make the market uninteresting for many potential buyers.

This is why we focus on optimizing the total utility of buyers subject to the constraint that the fixed costs of the allocated sellers can be recouped and the market is overall budget balanced.

3.2 Deriving a Value for Databases

To compute the allocation and payments of the sellers, we now derive the *induced* values of the buyers for the databases of the sellers. To this end, we first define the aggregate value of buyers for rows of database tables corresponding to their queries. Based on this aggregate value function, we can compute the positive externalities that different databases impose on buyers. Finally, we use these externalities to define buyers' values for different databases.

Given an allocation $a \in [0,1]^N$ and price per row p, the *market demand* for rows is $r^*(p,a) = \sum_{j=1}^{L} r_j^*(p,a)$. The market demand for money $m^*(p,a)$ is defined analogously, i.e., $m^*(p,a) = \sum_{j=1}^{L} m_j^*(p,a)$. We begin with the following definition:

Definition 4. *An **aggregate buyer** is a fictional buyer with an endowment $E = L \cdot e$ and the utility $U(m,r,a) = V(r,a) + m$. Here, $V(r,a)$ is an **aggregate value function**, i.e., a function that makes the solution of the consumption problem* (1) *for the aggregate buyer equal to the market demand $m^*(p,a), r^*(p,a)$.*

This means that the aggregate buyer is a fictional agent that acts in the same way as all buyers would act together when responding to the price p and the allocation a. The following proposition provides a way to compute the aggregate value function.

Proposition 3.1. *Given allocation $a \in [0,1]^N$, the aggregate value function is*

$$V(r,a) = \int_0^r \pi(z,a)dz, \tag{5}$$

where $\pi(z,a) = \max_{r^*(p',a)=z} p'$.[11]

Proof. From the Kuhn-Tucker conditions (Mas-Colell et al., 1995) for the aggregate consumer's problem (see Equation (1)), we have:

$$\frac{dV}{dr}(r^*, a) = p. \tag{6}$$

If $r^*(p,a)$ was a function, then the right hand side in Equation (6) would be the inverse demand function $r^{*-1}(p,a)$. However, $r^*(p,a)$ is a correspondence and, therefore, there can be many different prices p that support the solution $r^* = r^*(p,a)$. To resolve ambiguity, we let

$$\frac{dV}{dr}(r^*, a) = \max_{r^*(p',a)=r^*} p'. \tag{7}$$

Integrating Equation (7) and replacing $\max_{r^*(p',a)=z} p'$ with $\pi(z,a)$ we get $V(r,a) = \int_0^r \pi(z,a)dz$. Thus, solving Equation (6) with this aggregate value function gives us a solution $r^*(p,a)$, which is a market demand for rows. The demand of the aggregate buyer for money is then equal to $E - p \cdot r^*(p,a) = \sum_{j=1}^{L} e - p \sum_{j=1}^{L} r_j^*(p,a) = \sum_{j=1}^{L} m_j^*(p,a) = m^*(p,a)$, which is exactly the market demand for money. Thus, $V(r,a)$ is the aggregate value function. □

Now that we know how to compute the aggregate value function we can analyze some of its properties. We begin by showing that the aggregate value of buyers can only increase when more databases are allocated:

Proposition 3.2. *For all* $r \geq 0$, $\forall a \in [0,1]^N$, $\forall i = 1, ..., N$ *the following holds:*

$$\frac{\partial V}{\partial a_i}(r,a) \geq 0.$$

Proof. Consider a single buyer b_j. Let $\delta a = (0, ..., \delta a_i, ..., 0)$, with $\delta a_i \geq 0$. Then $\forall r$, $v_j(r, a + \delta a) = v_j(r,a) + \delta a_i \frac{\partial v_j}{\partial a_i}(r,a)$. Consequently, $\frac{\partial v_j}{\partial a_i}(r,a) = \frac{1}{\delta a_i}(v_j(r, a+\delta a) - v_j(r,a))$. From Assumptions 1 and 2 it follows that $v_j(r, a+\delta a) \geq v_j(r,a)$ for all $r \geq 0$, $\forall a, \forall j = 1, ..., L$. Therefore, $\frac{\partial v_j}{\partial a_i}(r,a) \geq 0$. This means that every buyer's value for r rows can only increase with the increase of a_i. Consequently, the aggregate value can also only increase. □

Example 1. *Assume that* $L = N = 2$ *and that* $e = 10$*; each seller produces a single database. Consider a setting where* s_1 *is allocated while* s_2 *is not, i.e.,* $a = (1,0)$*. Assume* b_j $(j = 1, 2)$ *submits a query against the database of* s_1 *and has the following value function for her data:* $v_1(r_1, a) = 4 \cdot \min\{r_1, 1\}$, $v_2(r_2, a) = 1 \cdot \min\{r_2, 2\}$*. Buyers solve their consumption problems, see Equation* (1)*. From the Kuhn-Tucker conditions we obtain the buyers' demands when*

11. This corresponds to the maximal *inverse demand function.*

the allocation is a:[12]

$$r_1^*(p,a) = \begin{cases} 1, & \text{if } p < 4 \\ [0,1], & \text{if } p = 4 \\ 0, & \text{otherwise} \end{cases} \qquad r_2^*(p,a) = \begin{cases} 2, & \text{if } p < 1 \\ [0,2], & \text{if } p = 1 \\ 0, & \text{otherwise.} \end{cases}$$

Consequently, the market demand is

$$r^*(p,a) = \begin{cases} 3, & \text{if } 0 < p < 1 \\ [1,3], & \text{if } p = 1 \\ 1, & \text{if } 1 < p < 4 \\ [0,1], & \text{if } p = 4 \\ 0, & \text{if } 4 < p. \end{cases} \tag{8}$$

It is easy to check that this market demand is equal to the demand of the aggregate buyer with an endowment of $E = 2e$ and utility $U(m,r,a) = V(r,a) + m$, where

$$V(r,a) = \begin{cases} 4r, & \text{if } 0 \le r \le 1 \\ 3 + r, & \text{if } 1 \le r \le 3 \\ 6, & \text{if } 3 \le r \end{cases}$$

is the aggregate value function. For example, here the aggregate value for the first row $(0 \le r \le 1)$ is equal to 4. Therefore, if $p > 4$, the aggregate buyer's demand must be 0. Equation (8) confirms this as $r^*(p,a) = 0$ for $p > 4$. Other cases are analogous.

Remember, that if a is an allocation, then $\lfloor a \rfloor_k$ stands for a similar allocation in which the database k is not allocated. We now define the *externality* imposed by a database k on all buyers as follows:

Definition 5. *For a given allocation a and a posted price p, the externality imposed by the database k is*

$$ext_k(a,p) = V(r^*(p,a),a) - V(r^*(p,\lfloor a \rfloor_k),\lfloor a \rfloor_k). \tag{9}$$

The externality reflects how much additional value the database k brings to all buyers. Note that this quantity could be zero if the database k is not allocated in a. To define the value of buyers for a database k in a consistent way, we split the aggregate value achieved by all buyers proportionally to $ext_k(a,p)$. Formally:

Definition 6. *Given allocation $a \in \{0,1\}^N$ and price p, the **induced value $W_k(a,p)$ of the aggregate buyer for the database k** is the share of the aggregate value that is proportional to the externality that database k imposes on the aggregate buyer, i.e.,*

$$W_k(a,p) = \frac{ext_k(a,p)}{\sum_{\ell=1}^{D} ext_\ell(a,p)} V(r^*(p,a),a). \tag{10}$$

12. The result is quite intuitive. Indeed, b_j is not willing to buy query answers as long as the price p per answer is larger than his marginal value for the answer (which is equal to 4 for b_1 and 1 for b_2). As soon as the price is smaller than the marginal value, b_1 and b_2 are willing to buy up to one and two query answers respectively.

Observe that $W_k(a, p)$ depends on the posted price p and on the whole allocation a. Thus, indirectly it depends on allocations of all other databases. The dependency on allocation a reflects the important fact that buyers may have combinatorial valuations, i.e., they can value database k higher if a complementary database is also available.

The fact that $W_k(a, p)$ depends on price p has the following intuition: If price p is too high such that no one can afford to buy a single row, the externality imposed by any database is zero. If prices are low, externalities become positive.

We impose an additional assumption: all allocated databases are (weak) complements for the aggregate buyer. Thus, his valuation of databases is supermodular (Chambers & Echenique, 2009). Intuitively, supermodularity can be described as follows. Consider any two databases ℓ, k. Supermodularity says that the induced value of database k can only increase as the allocation probability of the other database ℓ increases. Formally:

Assumption 3 (**Weak Complementarity**). *For any two databases ℓ and k, for any $a \in \{0, 1\}^N$, $\forall p \geq 0$ the following inequality holds:*

$$W_k(a, p) \geq W_k(\lfloor a \rfloor_\ell, p). \tag{11}$$

Remark 3.1. *This assumption is quite intuitive: If the database ℓ complements another database k, an increase in allocation probability of ℓ leads to an increase in the induced value of database k for the aggregate buyer. Imposing this assumption on the aggregate buyer is natural: given that this buyer can be considered as the whole population, it makes sense to allocate those databases that are complementary. Note that, for individual buyers different databases can still be either complements or substitutes.*

While Assumption 3 is intuitive, it does not follow directly from Assumption 1 and Assumption 2. Indeed, one could construct an example where Assumptions 1 and 2 are satisfied but Assumption 3 is not. To eliminate these corner cases we state the Assumption 3 explicitly.

Example 2. *We follow the setup of Example 1. The externality that s_1 imposes on buyers when the allocation is $a = (1, 0)$ and the posted price is p is $ext_1(a, p) = V(r^*(p, a), a) - V(r^*(p, \lfloor a \rfloor_1), \lfloor a \rfloor_1)$, where $\lfloor a \rfloor_1 = (0, 0)$. Here, the aggregate value when the allocation is $\lfloor a \rfloor_1$ is $V(r, \lfloor a \rfloor_1) = 0$ as no queries can be answered. Similarly, the externality that s_2 imposes on buyers under allocation a is 0 (as s_2 is not allocated). Thus, for example, if $p = 1 - \epsilon$, then the market demand $r^*(p, a) = 3$ and the induced value of the database of s_1 is $W_1(p, a) = \frac{6}{6+0} \cdot 6 = 6$. Example 3 extends the current example for the case when all databases are allocated.*

3.3 Designing the Reverse Auction

Now that we have defined the induced values for different databases, $W_k(a, p)$, we can define an appropriate auction $\mathcal{A}$ that maximizes the total utility of buyers, subject to the constraint that the fixed costs of allocated sellers are recouped. In this auction, the market platform computes the allocation a and payments t based on the costs reported by sellers. We design this auction in a way that is similar to the optimal auction of Myerson (1981).

First, observe that the expected total utility of buyers is equal to the difference between the expected value that the aggregate buyer can achieve under the allocation $g(c)$ and the total payment that the buyers must make to the sellers. Formally,

$$\begin{aligned}\mathbb{E}_f[U(g(c),h)] =& \mathbb{E}_f\Big[V(r^*(p,g(c)),g(c)) - \sum_{i=1}^{N} h_i(c)\Big] \\ =& \mathbb{E}_f\Big[\sum_{k=1}^{D} \mathbb{E}_{g(c)}\Big[W_k(a,p)\Big] - \sum_{i=1}^{N} h_i(c)\Big]. \end{aligned} \tag{12}$$

Here, $\mathbb{E}_{g(c)}\Big[W_k(a,p)\Big]$ is the expected induced value of the database k with the probabilistic allocation of sellers $g(c)$ (see the formal definition of this term in Appendix B). The market design problem now is to find an auction $\mathcal{A} = \langle g, h\rangle$ that maximizes $\mathbb{E}_f[U(g(c),h)]$ subject to BNIC, IR and the following constraints:

$$\sum_{i\in i(k)} g_i(c) \leq 1 \qquad \forall k = 1,...,D \;\; \forall c \in [0,1]^N, \tag{13}$$

$$g_i(c) \geq 0 \qquad \forall i = 1,...,N \;\; \forall c \in [0,1]^N. \tag{14}$$

Constraints (13) and (14) ensure that each database is allocated at most once and that the allocation probabilities are non-negative.

We let $\phi_i(c_i) = c_i + \frac{F_i(c_i)}{f_i(c_i)}$ denote the *virtual cost* of seller s_i. We assume that the distributions of costs of sellers f_i are regular, i.e., $\phi_i(c_i)$ are monotone and strictly increasing (Myerson, 1981). Let us also define the *virtual surplus* as $\widetilde{S}(a) = \sum_{k=1}^{D} W_k(a,p) - \sum_{i=1}^{N} \phi_i(c_i)a_i$, $a \in \{0,1\}^N$. We let $\lfloor \tilde{S} \rfloor_i(a)$ denote the virtual surplus achieved by all agents apart of s_i, i.e., $\lfloor \tilde{S} \rfloor_i(a) = \sum_{k=1}^{D} W_k(a,p) - \sum_{j=1,j\neq i}^{N} \phi_j(c_j)a_j$. Now, we are ready to present the optimal reverse auction.

Buyer-Optimal Reverse Auction (BORA)

Allocation rule: $a^* \in \arg\max_{a\in\{0,1\}^N} \widetilde{S}(a)$; use random tie breaking in the case of ties.

Payment rule: For each seller s_i:

If $a_i = 1$, then

$$t_i = \phi_i^{-1}\Big(\lfloor \widetilde{S} \rfloor_i(a^*) - \widetilde{S}(\lfloor a \rfloor_i^*)\Big). \tag{15}$$

If $a_i = 0$, then $t_i = 0$.

In words, the allocation rule says that the auction allocates sellers in a way that maximizes the virtual surplus. We break ties randomly. Informally, the payment of the allocated agent is computed in a similar way as VCG payments, where agents report their virtual costs instead of their true costs. To better understand the intuition behind the payment rule, let us consider several special cases.

One database, one seller. Let us first consider the setting with a single seller, i.e., $N = 1$, and consequently, $D = 1$. In this case, the seller is allocated whenever her virtual cost is smaller than the induced value of buyers for her database, i.e., $\phi_1(c_1) \leq W_1(a, p)$. From Equation (15) it follows that the payment to the seller must be equal to $t_1 = \phi_1^{-1}(W_1(a, p))$. This payment is similar to the Myerson (1981) *reserve payment*. In words, the reserve payment is equal to the cost that the seller would have had to make her virtual cost to be equal to the induced value of the aggregate buyer for her database.

One database, multiple sellers. Let us now assume that there are multiple sellers that can produce the same database, i.e., $D = 1$ and $N > 1$. In this case, the seller with the smallest virtual cost is allocated as long as her virtual cost is smaller than the induced value of buyers for her database. W.l.o.g., let us assume that $\phi_1(c_1)$ is the smallest virtual cost and $\phi_2(c_2)$ is the second smallest virtual cost. Then, the payment of the allocated seller is equal to the minimum of $\phi_1^{-1}(\phi_2(c_2))$ and $\phi_1^{-1}(W_1(a, p))$, i.e., $t_1 = \min\{\phi_1^{-1}(\phi_2(c_2)), \phi_1^{-1}(W_1(a, p))\}$. In other words, the payment of the allocated agent is computed as the minimum of the reserve payment and the *critical value*. Here, the critical value is defined similarly to (Nisan et al., 2007), i.e., it is equal to the largest cost that the seller could have reported while still being allocated, i.e., $\phi_i^{-1}\Big(\min_{j \in i(k) \setminus i} \phi_j(c_j)\Big)$.

Two databases, two sellers. Consider the setting with two distinct databases, each produced by a single seller. Assume further that $\phi_1(c_1) > W_1(a, p)$ for $a = (1, 1)$ and for all p. In contrast to the setting with a single database and a single seller discussed above, in this case, one can happen that the database 1 is allocated (i.e., despite of the fact that its virtual cost is larger than the induced value of the respective database). Such a situation is possible, for example, when the database 1 has a very strong complementary effect on the database 2. Thus, the presence of the database 1 can increase the induced value of the second database, $W_2(a, p)$, as this induced value depends on the *whole* allocation a. As a result this may lead to a higher virtual surplus. Therefore, it may be optimal to allocate both databases. Example 4 in Appendix A illustrates this case.

Multiple databases, multiple sellers. In this most general case, the intuition behind the payment rule (15) mimics the intuition of the standard VCG mechanism. Indeed, the first two summand in the Equation (15), $(\phi_i(c_i) + \widetilde{S}(a^*))$, correspond to the total virtual surplus achieved by all sellers apart of s_i at the optimal allocation a^*. The last summand, $\widetilde{S}(\lfloor a \rfloor_i^*)$, corresponds to the optimal virtual surplus achieved in a similar setting but without the seller s_i being present. Thus, the argument of the inverse virtual cost function can be interpreted as a *virtual externality* imposed by the seller s_i.

Observe also that in the general case, the objective of maximizing the virtual surplus in the allocation rule of the BORA auction is non-linear. This follows from the fact that the induced values of databases need not depend linearly on different allocations. This poses a number of computational challenges that we will address in Section 3.5. Appendix A presents a number of worked examples that illustrate the BORA auction.

Theorem 3.3. *If the distributions f_i are regular for all $i = 1, ..., N$, and the databases are weakly complementary for the aggregate buyer, then the BORA auction maximizes buyers utilities and satisfies constraints* (13), (14), *BNIC and IR.*

Proof. First, let $\mathbb{E}_{f_{-i}}[g_i(c_i, c_{-i})]$ be an *ex-interim* allocation probability of s_i. As Myerson (1981) showed, BNIC, IR and constraints (13) and (14) imply monotonicity of ex-interim allocation (see Lemma B.1 in Appendix B). We use this result to prove the following lemma:

Lemma 3.4. *Consider the allocation rule* $g : \mathbb{R}^N_{\geq 0} \to [0,1]^N$ *that maximizes*

$$\mathbb{E}_f\Big[\sum_{k=1}^{D}\Big(\mathbb{E}_{g(c)}\Big[W_k(a,p)\Big] - \sum_{i\in i(k)}\Big(c_i + \frac{F_i(c_i)}{f_i(c_i)}\Big)g_i(c)\Big)\Big] \tag{16}$$

subject to monotonicity of the ex-interim allocation and constraints (13) *and* (14)*. Further, consider the payment rule* $h_i(c) = g_i(c)c_i + \int_{c_i}^{\beta_i} g_i(\hat{c}_i, c_{-i})d\hat{c}_i$ *for every* $i = 1, ..., N$, $\forall c$*. Then,* $\mathcal{A} = \langle g, h\rangle$ *maximizes buyers' utilities under the constraints* (13) *and* (14)*, BNIC and IR.*
Proof. The proof is presented in Appendix B. □

Now we would like to show that if a database ℓ is allocated with a positive probability, then this probability must be equal to 1. To achieve this, we first present the following proposition that shows that constraints (13) are binding when databases are complements:

Lemma 3.5. *Let* $g^*(c)$ *be a solution of* (16)*. Then, if for a database* ℓ *there exists a seller* s_i *with* $i \in i(\ell)$ *such that* $g_i^*(c) > 0$*, then* $\sum_{i\in i(\ell)} g_i^*(c) = 1$.

Proof. See Appendix B for the proof. □

From Lemma 3.5 and Lemma B.2 (see Appendix B), it follows that, there must exist a deterministic allocation $g^*(c)$ that maximizes Equation (16). This narrows down the search space of optimal mechanisms to only the *deterministic* ones. Consider a mechanism that for any reported cost profile c maximizes the virtual surplus $\widetilde{S}(a) = \sum_{k=1}^{D} W_k(a,p) - \sum_{k=1}^{D}\sum_{i\in i(k)} \phi_i(c_i)a_i$, where $a \in \{0,1\}^N$. This allocation also maximizes Equation (16). Remember that the distributions f_i are regular.[13] Thus, $\phi_i(c_i)$ must be monotone. Consequently, the ex-interim allocation is also monotone.

From Lemma 3.4 it follows that in order guarantee BNIC, the payments of sellers must satisfy

$$h_i(c) = g_i^*(c)c_i + \int_{c_i}^{\beta_i} g_i^*(\hat{c}_i, c_{-i})d\hat{c}_i \tag{17}$$

for every $i = 1, ..., N$, $\forall c$. If a seller s_j is not allocated (i.e., $a_j = 0$), then from monotonicity of the ex-interim allocation (see Lemma B.1) it follows that s_j is not allocated for any other cost $q_j \geq c_j$. Consequently, $h_j(c) = 0$. If a seller s_j is allocated (i.e., $a_j = 1$), then Equation (17) can be simplified as follows

$$h_j(c) = c_j + \int_{c_j}^{\zeta_j} d\hat{c}_j = c_j + \zeta_j - c_j = \zeta_j, \tag{18}$$

where

$$\zeta_j = \sup\{c | \phi_j(c) \leq \phi_i(c_i) \quad \forall i \in i(k)\backslash j \ \texttt{and}\ \phi_j(c) \leq \phi_j(c_j) + \widetilde{S}(a^*) - \widetilde{S}(\lfloor a \rfloor_j^*)\}. \tag{19}$$

13. For irregular distributions we could use *ironing* in a similar way as in (Myerson, 1981).

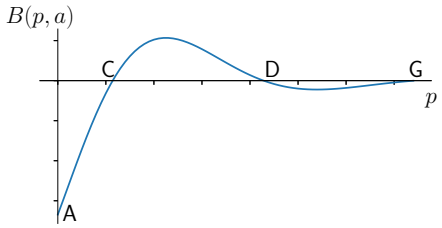

Figure 4: Example of the dependency of the budget surplus $B(p, a)$ on posted price p. Point C corresponds to the minimal p that satisfies overall budget balance.

Now, let us exclude the seller s_j who produces a database k from the mechanism and consider two scenarios. In the first scenario, the resulting allocation of databases stays the same, i.e., the database k is still produced but perhaps by a different seller $i \in i(k)$. In this case, it must hold that $\phi_i(c_i) = \min\{\phi_q(c_q), q \in i(k)\}$. Thus, the second inequality in (19) implies the first one. In the second scenario, the database k is not allocated anymore. Thus, it must be that $\phi_i(c_i) - \phi_j(c_j) \geq \tilde{S}(a^*) - \tilde{S}(\lfloor a \rfloor_j^*)$, where $\phi_i(c_i) = \min\{\phi_q(c_q), q \in i(k)\}$. Equivalently, $\phi_i(c_i) \geq \phi_j(c_j) + \tilde{S}(a^*) - \tilde{S}(\lfloor a \rfloor_j^*)$ and therefore, the first inequality in (19) is again implied by the second one. Therefore, from the monotonicity of $\phi_j(c_j)$ it follows that the payment of any seller s_j can now be rewritten as follows:

$$h_j(c) = \phi_j^{-1}\Big(\phi_j(c_j) + \widetilde{S}(a^*) - \widetilde{S}(\lfloor a \rfloor_j^*)\Big). \tag{20}$$

□

3.4 The Overall Market Mechanism

The reverse auction designed in the previous section does not guarantee that the market mechanism is budget balanced. Instead, it assumes that the market platform can always pay the sellers. To guarantee that the market is budget balanced, the market mechanism needs to set a posted price p, such that the total amount of money collected from the buyers, $\sum_{j=1}^{L}(e - m_j^*(p, a))$, is equal to the total payment $\sum_{i=1}^{N} t_i$ received by the sellers.

Consider the *budget surplus*[14] achieved when the allocation is a and the price is p:

$$B(p, a) = \sum_{j=1}^{L}(e - m_j^*(p, a)) - \sum_{i=1}^{N} t_i. \tag{21}$$

Observe that the total payment to be accrued to sellers depends on allocation a and on price p. To see this, notice that if $p = 0$, buyers can gain a lot of value from submitting queries against all databases (for free). Thus, the induced value of any database k, $W_k(a, p)$, is large, and it is likely that the database is allocated. As a result, the second term in (21) is positive. At the same time, buyers do not pay anything and, consequently, the first term in Equation (21) is zero. This means that if $p = 0$, then $B(p, a) < 0$. This case is illustrated by point A in Figure 4. A similar argument works for a situation in which $p = \infty$. In this case, both terms in $B(p, a)$ are zero, which corresponds to the trivial equilibrium when no

14. In microeconomic literature, this quantity is also often called the *excess demand* for money (Mas-Colell et al., 1995).

sellers are allocated. Point G in Figure 4 illustrates this scenario. As can be seen in Figure 4, the non-trivial equilibrium prices that satisfy the budget balance constraint correspond to points C and D. In general, there may be multiple solutions for which $B(p, a) = 0$ (such as points G, C and D in Figure 4). Consequently, there may be many different posted prices that guarantee budget balance. However, we aim to find the smallest such price, as it would deliver the largest total utility to buyers.

Data: $F_i(c_i)$, $i = 1, ..., N$; $F_{v'(a)}$, $F_{\overline{r}(a)}$ for all $a \in \{0, 1\}^N$
Result: Allocation a, payments t, posted price p

1 $\delta \leftarrow 0.01$ // Step size
2 $\iota \leftarrow 0$
3 $p(\iota) = 0$
4 $a(\iota) \leftarrow (a_1, ..., a_N)$, s.t., $\forall k \leq D$ holds $\sum_{i \in i(k)} a_i = 1$
5 ask sellers to report $\hat{c} = (\hat{c}_1, ..., \hat{c}_N)$
6 **repeat**
7 compute $W_k(a(\iota), p(\iota))$ for all $k \leq D$ // See Definition 6
8 set up $\mathcal{A} = \langle g, h \rangle$ parametrized by $W_k(a(\iota), p(\iota))$ // See BORA
9 solve $\mathcal{A}$, i.e., compute $a \leftarrow g(\hat{c})$ and $t \leftarrow h(\hat{c})$ // See BORA
10 compute buyers' demand for money $m_j^*(p(\iota), a(\iota))$
11 $B(p(\iota), a(\iota)) \leftarrow \sum_{j=1}^{L} \left(e - m_j^*(p(\iota), a(\iota))\right) - \sum_{i=1}^{N} t_i$ // Budget surplus
12 $p(\iota + 1) \leftarrow p(\iota) - B(p(\iota), a(\iota)) \cdot \delta$ // Price update
13 $\iota \leftarrow \iota + 1$
14 **until** $|B(p(\iota), a(\iota))| \leq \epsilon$;
15 $p \leftarrow p(\iota)$
16 **return** a, t, p

Algorithm 1: Fixed-point iteration for computation of the allocation and the price.

To find a solution, we adopt an idea similar to the Tatonnement process (Cheng & Wellman, 1998). More concretely, in Algorithm 1, we design an iterative procedure that updates the price, allocation and payments of sellers at every iteration ι (see line 2 of Algorithm 1). We begin from an initial price $p_0 = 0$ that corresponds to a situation when rows of database tables corresponding to buyers' queries are free. At this step we also assume that all databases are allocated (lines 3-4 of Algorithm 1). In lines 6-14 of Algorithm 1, we then perform a fixed-point iteration by increasing the posted price $p(\iota)$ as a function of the iteration ι, as well as adjusting allocation probabilities $a(\iota)$. At every iteration ι, we evaluate $W_k(a(\iota), p(\iota))$ and compute the tentative allocation a and payments t of the auction $\mathcal{A}$ (lines 7-9). Observe, that given the price $p(\iota)$ we can also compute the buyers' demand for money $m_j^*(p(\iota), a(\iota))$ (see line 10). Consequently, the amount of money that buyers are willing to pay for their queries at iteration ι is $\sum_{j=1}^{L} \left(e - m_j^*(p(\iota), a(\iota))\right)$. Thus, in line 11 of Algorithm 1, we now can compute the budget surplus, $B(p(\iota), a(\iota))$, at this iteration. As long as $B(p(\iota), a(\iota))$ is negative, we increase the price $p(\iota)$. We stop the algorithm when the budget surplus is smaller than the chosen tolerance threshold (see line 14 of Algorithm 1).

Remember that the sellers are asked to report their costs only once. Precisely, this happens at time $\tau = 1$ of our temporal model (see Figure 3). We assume that the sellers understand Algorithm 1 and the rules of the BORA auction. Thus, they can make a truthful report $\hat{c} = c$ (see line 5 of Algorithm 1).

Note also that Algorithm 1 does not require any interaction with the buyers. Instead, it is considered a heuristic procedure for computing an equilibrium price and allocation. Thus, this algorithm must be executed at time period $\tau = 2$ of our temporal model (see Figure 3), in other words, *before* actual buyers arrive to the market. This implies that the iterative nature of the algorithm does not change the incentives of the buyers to behave truthfully and the truthfulness of the overall market follows immediately from the truthfulness of the BORA auction.

Second, observe that a non-trivial equilibrium does not always exist. Consequently, our algorithm may return a null allocation and zero payments. Consider, for example, about a domain with a single seller with a high fixed cost and assume that there is a single buyer with a very small marginal value and a small value threshold. In this case, it is not possible to compensate the seller for producing her database. This result, however, does not constitute a failure of our market design: Indeed, if the society does not value the data highly enough, then the data should not be produced in the first place. In other words, we are aiming at designing a market that incentivizes data providers to produce *useful* data rather than *any* data. [15]

Finally, note that, even though we designed our market with the goal of optimizing the buyers' surplus, it is not possible to provide any meaningful lower bound on the share of the surplus obtained by buyers in general. The following proposition states this result formally:

Proposition 3.6. *The share of the buyers' surplus achieved by Algorithm 1 auction is lower bounded by zero.*

Proof. To prove the statement we construct a corner case where all sellers have zero costs and face no competition for producing their databases. At the same time, only joining *all* databases brings value to every buyer, while joining any other combination of databases has zero value. In this case, we can show that the buyers' surplus is zero (and it can't be negative as buyers can always decide not to participate in the trade). The full proof is provided in Appendix B. □

The corner case used to prove Proposition 3.6 is obviously pathological, and we would not expect such cases in practice. To study how much surplus buyers can get in more realistic settings, we have performed a number of computational experiments (see Section 4).

3.5 Winner Determination via Mixed-Integer Allocation Programming

In this section, we discuss computational challenges that arise in practical implementation of our proposed BORA auction. First, remember that buyers are indifferent about the

15. Remember, that our choice of the initial price $p_0 = 0$ follows the idea that we want to find an equilibrium with the largest surplus for buyers. Clearly, if the initial price was too high, then the trivial equilibrium in which nobody is allocated could be reached immediately. Technically, we could also launch Algorithm 1 from several starting points. However, our experiments show that even starting with $p = 0$, we can obtain high levels of surplus. Example 5 in Appendix A illustrates our approach.

identities of sellers. Thus, it follows that the induced values of databases $W_k(a,p)$ are constant for any allocation of sellers as long as the allocation of the respective databases stay the same. As the number of databases D is typically much lower than the number of sellers N, we can now compute the induced values of databases for every possible deterministic allocation $\alpha \in \{0,1\}^D$ of databases rather than for every possible deterministic allocation $a \in \{0,1\}^N$ of sellers. With a slight abuse of notation we let $W_k(\alpha,p)$ be the induced value of the database k when the *allocation of databases* is α.

Further, remember that the winner determination problem in the BORA auction is non-linear. In order to linearize it, we can pre-compute the induced values of databases for every possible deterministic allocation of databases. We then include these pre-computed values into the objective function with auxiliary binary optimization variables indicating whether a particular deterministic allocation of databases is chosen. This idea is illustrated with the following linearized mixed integer program:

$$\max_{\substack{a_i, i=1,...,N \\ z_\alpha, \alpha \in \{0,1\}^D}} \sum_{\alpha \in \{0,1\}^D} \left[z_\alpha \cdot \sum_{k=1}^{D} W_k(\alpha,p) \right] - \sum_{i=1}^{N} \phi_i(c_i) a_i \tag{22}$$

$$\texttt{s.t.} \quad z_\alpha \leq \sum_{j \in i(k)} a_j \quad \forall \alpha \in \{0,1\}^D \ \ \forall k = 1,...,D \ \ \texttt{s.t.} \quad \alpha(k) = 1 \tag{23}$$

$$z_\alpha \leq \sum_{j \in i(k)} (1 - a_j) \quad \forall \alpha \in \{0,1\}^D \ \ \forall k = 1,...,D \ \ \texttt{s.t.} \quad \alpha(k) = 0 \tag{24}$$

$$a_i \in \{0,1\} \quad \forall i = 1,...,N \tag{25}$$

$$z_\alpha \in \{0,1\} \quad \forall \alpha \in \{0,1\}^D. \tag{26}$$

Here, binary optimization variables a_i represent the allocation decisions of the BORA auction regarding sellers s_i, $i = 1,...,N$. Further, the auxiliary optimization variable z_α is equal to 1 if the deterministic allocation of databases is $\alpha \in \{0,1\}^D$. Here, constraints (23) and (24) build a bridge between allocation decisions regarding different sellers and the chosen deterministic allocation of databases produced by these sellers. In particular, constraints (23) guarantee that the deterministic allocation α in which the database k is allocated ($\alpha(k) = 1$) is not feasible ($z_\alpha = 0$) if none of the sellers producing database k are allocated. Similarly, the constraint (24) sets $z_\alpha = 0$ if at least one seller producing the database k is allocated even though the database k should not be allocated ($\alpha(k) = 0$).

Looking into the objective function (22), we see that constraints (23) and (24) guarantee that there is only a single term of the total induced value of databases that gets activated for a specific allocation of databases and sellers. In particular, if the allocation of databases is α and the allocation of sellers a satisfies constraints (23) and (24), then the objective function value is $\sum_{k=1}^{D} W_k(\alpha,p) - \sum_{i=1}^{N} \phi_i(c_i) a_i$.

Observe, that such a linearization of the allocation problem of the BORA auction comes at a high cost. Indeed, in order to achieve the linear formulation, we have introduced a number of auxiliary optimization variables z_α, and this number grows exponentially in the number of databases D. However, the high complexity of our approach seems to be unavoidable and follows directly from the combinatorial preferences of buyers for different

allocations of databases.[16] Example 4 in Appendix A illustrates the advantages of using the combinatorial design of the auction to achieve higher buyer surplus compared to a non-combinatorial design.

4. Experiments

To study the economic properties of our market design, we carry out a set of computational experiments. These experiments allow to quantify a number of practically important economic parameters (such as surplus distribution, monopoly power of unique data providers etc.) and to check how robust are some of the assumptions we have made at scale. To do this, we first implement a simulation set up to generate buyers and sellers according to the model described in Section 2.1. We then run our market mechanism based on Algorithm 1 and measure the social welfare, as well as the share of the social welfare obtained by buyers and sellers, respectively. Note that, at first, we perform all of our experiments in a small "stylized" setting to gain a detailed understanding of the behavior of our market mechanism. To study the scalability of our approach, we then perform a number of experiments in a medium-sized domain, in which we increase the number of databases, sellers and buyers. Even more realistic, large-scale market simulations would require a thorough examination of buyers' preferences for data in a particular domain. We defer such large-scale simulations to future work.

4.1 Small Experimental Set-up

We simulate a setting with $N = 3$ sellers and $D = 2$ databases. We assume that seller s_1 can produce database 1, and both sellers s_2 and s_3 can produce database 2. We assume that costs c_i of all sellers are i.i.d., $c_i \sim U[0, 20]$, $i = 1, 2, 3$. This models a scenario in which seller s_1 is a unique producer of database 1 (i.e., s_1 has a monopolistic ownership of her data), while sellers s_2 and s_3 are competing to produce database 2. Because the BORA auction is BNIC (Theorem 3.3) we assume that sellers report their costs truthfully.

We vary the number of buyers L from 1 to 128 while keeping the number of sellers and databases fixed. Each buyer has an initial endowment of money $e = 10$. For simplicity, we assume that the marginal value v'_j and the threshold $\overline{r}_j$ of any buyer b_j only depends on the number of allocated databases but not on the identities of these databases. If there are no databases allocated, the marginal value and the threshold of each buyer are equal to zero. If there is exactly one allocated database, then the marginal value of each buyer is drawn from $U[0, 2]$. The threshold, in this case, is drawn from a discrete uniform distribution $U\{0, 5\}$. Finally, if there are exactly two databases allocated then the marginal value v_j is incremented by a random variable drawn from $U[0, 2]$, while $\overline{r}_j$ is incremented by a random variable drawn from $U\{0, 5\}$.

We call the auction instances with the same number of buyers a *setting*. For each setting (with 1, 2, 4, ... buyers) we generate 10 random instances as described above. This allows us to estimate the mean values and confidence intervals for every setting.

16. In Section 4.2, we provide a further discussion of how this complexity can be reduced by partitioning available databases into loosely connected sets, and running the market mechanism for each set of such a partitioning.

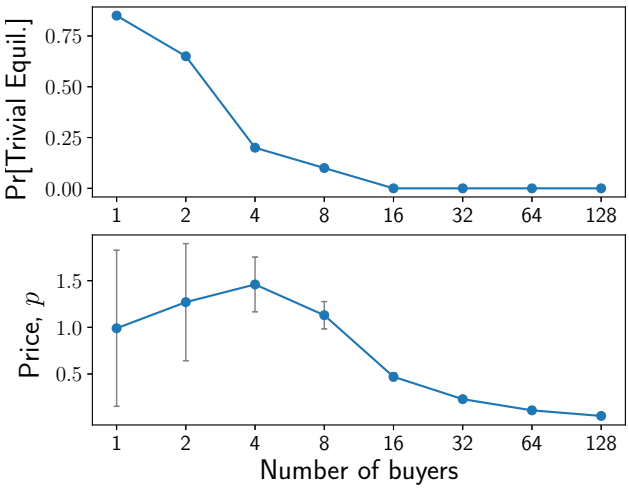

Figure 5: The top graph shows the probability that the market is in a trivial equilibrium. We vary the number of buyers from 1 to 128. The bottom graph shows the dependency of price p on the number of buyers, L. Error bars indicate confidence intervals at 0.05 significance level.

4.1.1 Probability of "No Trade"

Consider Figure 5 (top), which shows the probability that the market mechanism only finds the *trivial equilibrium* ("no trade"). We see that when L is small (≤ 8), then it is likely that there are not enough buyers to cover the fixed costs of sellers, whatever the price p. If $L \geq 8$, then both databases are always allocated. In this case, seller s_1 produces database 1 and either s_2 or s_3 produces database 2.

4.1.2 Posted Price

Consider Figure 5 (bottom), which shows the dependency of the posted price (in a non-trivial equilibrium) on the number of buyers L. As L grows, the amount of money that must be collected from each buyer to achieve BB decreases. Consequently, the price p also decreases. As the number of buyers increases, the marginal effect of every additional buyer on the price decreases (the price curve becomes less steep). This is expected, as the impact of an individual buyer on the aggregate demand decreases as L increases, which leads to more and more"price-taking behavior".

4.1.3 Efficiency and Social Welfare

We now study the most important question: How *efficient* is our market mechanism? To this end, consider Figure 6. In the graph in the center, we observe that the absolute value of social welfare grows linearly as the number of buyers increases (an exponential trend in the logarithmic scale). This is what we would expect. Now, consider Figure 6 (top),

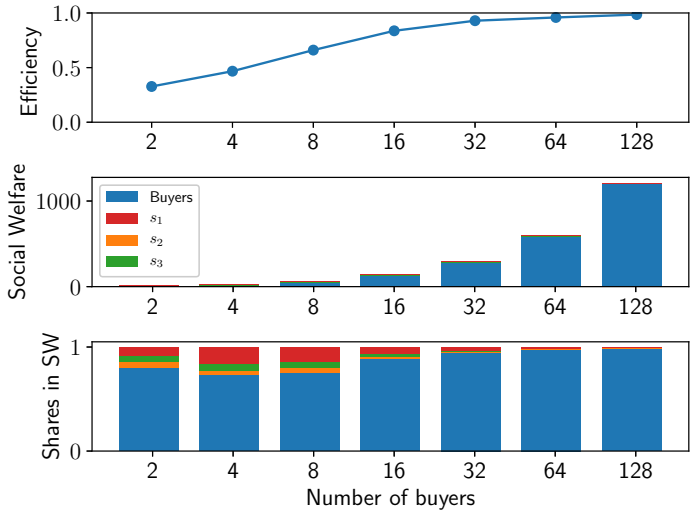

Figure 6: The top graph shows efficiency reached in a non-trivial equilibrium. The middle graph shows social welfare separated by buyers and allocated sellers. The bottom graph shows the relative distribution of the achieved welfare. Small differences in the mean surplus of sellers s_2 and s_3 are not statistically significant at the 0.05 significance level.

which shows the efficiency of our market mechanism.[17] For illustration reasons we omit the efficiency measurement that corresponds to $L = 1$. [18] If both databases are allocated ($L > 32$), the efficiency is 95%. As L increases further, the efficiency stabilizes. This is due to the fact that the posted price p becomes more flat (see Figure 5 (bottom)), such that there is a constant fraction of buyers with a marginal value per row smaller than the posted price p. These buyers do not buy anything, causing the efficiency loss. Thus, as the change in p gets smaller for larger number of buyers, the efficiency also becomes nearly constant.

4.1.4 Shares of Social Welfare

Consider Figure 6 (bottom), which shows how social welfare is distributed between buyers and sellers. Observe that seller s_1's share increases as the number of buyers increases (as long as $L \leq 8$), while the shares of the other sellers decrease (see also Table 1 for the absolute values and standard errors). To understand this result, remember that seller s_1 is a monopolist, i.e., she faces no competition for her database. Thus, her payment is solely determined by the reserve price set by the auction, which only depends on the value that database 1 is expected to generate for *all buyers*. As L increases, this value naturally increases, which means that the seller receives a larger payment. However, this payment is bounded by the upper bound of the support of the distribution of sellers' costs (which is 20 in this set up). This is why as soon as $L \geq 16$, the share of s_1 can only decrease. In

17. Efficiency is defined in the standard way, as the fraction of the social welfare achieved by our mechanism and the social welfare of an *optimal* (omniscient) mechanism which can disregard incentive constraints.
18. The reason for that is that in this case our market reaches almost 100% efficiency. This is due to the fact that when the number of buyers is small, the market stays in the "no trade" equilibrium most of the time (see Figure 5 (top)). Even when there is trade, only one database gets allocated, which makes achieving high efficiency easier.

Surplus	Number of buyers						
	2	4	8	16	32	64	128
Buyers	7.20 (1.28)	17.66 (2.25)	45.00 (6.17)	128.02 (8.23)	280.82 (6.18)	583 (7.97)	1198.23 (15.65)
Seller s_1	1.17 (0.62)	4.39 (1.46)	8.61 (1.40)	9.57 (1.26)	9.57 (1.26)	9.57 (1.26)	9.57 (1.26)
Seller s_2	0.54 (0.35)	0.80 (0.62)	3.40 (1.16)	4.06 (1.33)	4.06 (1.33)	4.06 (1.33)	4.06 (1.33)
Seller s_3	0.59 (0.41)	1.74 (0.67)	3.26 (0.87)	3.43 (0.97)	3.43 (0.97)	3.43 (0.97)	3.43 (0.97)

Table 1: Mean values and standard errors for the surplus of buyers and sellers for the small setup reached in a non-trivial equilibrium.

contrast, s_3's payment is constant for $L \geq 20$ and depends only on the virtual cost of her competitor s_2.[19] Finally, there is no significant evidence that sellers s_2 and s_3 get different surplus in the market at the 0.05 significance level, as is expected due to their symmetry.

4.2 Medium Experimental Set-up.

To study the scalability of our approach, we perform a number of computational experiments with more realistic numbers of sellers, databases and buyers. In practice, the exact numbers would be domain specific. We expect them to be on the order of dozens to hundreds for sellers and thousands for buyers. Some evidence for these numbers is derived from an examination of existing data marketplaces. For example, a recent report on the Oracle proprietary digital data marketplace for selling marketing data presents a domain with more than 200 data providers and 200 customers across multiple industries (Blue Kai, Inc., 2011). Assuming that a typical query against real-world databases joins only two to four of these databases, we can split buyers into different groups based on the databases they are most often interested in. [20] Thus, in our medium-sized experiments, we vary the number of databases D from 2 to 10. While some databases can be produced by unique data providers (monopolistic data ownership), others can be produced by many different data providers. This allows us to vary the number of data providers N from 2 to 100. Similarly to the small set-up, we generate sellers' costs from a uniform distribution $U[0, 20]$.

In this set-up, we vary the number of buyers L from 8 to 1024. We use the same simple value model for buyers as in the small set-up. Specifically, we assume that the allocation of an additional database has two effects on buyers. First, it leads to an increase in the marginal

19. When $L \leq 8$, then the reserve payment is smaller than the virtual cost of s_2. Then s_3's payment sometimes depends on the reserve payment.

20. In this case, buyers within a group are assumed to submit most of their queries against 5 to 15 databases and very few queries against the remaining databases. In other words, we assume that the overall market can be partitioned into several sets of databases with loose pairwise connections between these sets. Such partitioning reduces the number of databases and buyers in each set, which therefore leads to a lower computational burden. Therefore, for our experiments, we assume that such a partitioning can be carried out effectively using a certain clustering technique, and we can run our market mechanism for each set of the partitioning. The design of the exact clustering procedure is beyond the scope of this work.

value v_j of a buyer by a random amount drawn from $U[0, 2]$. Second, it causes an increase in the threshold $\overline{r}_j$ by a random number drawn from $U\{0, 5\}$. The main complication of the medium-sized set-up compared to the small set-up is of a computational nature: Due to the combinatorial structure of the problem, we must now specify the value function of every buyer for all 2^D possible deterministic allocations of databases. As previously discussed, in practice, many buyers submit their queries against only a small subset of available databases. This means that, on average, the marginal change of a buyer's value due to the allocation of an additional database gets smaller as the number of allocated databases grows. As a result, the computational problem of estimating the typical value function of a buyer may be somewhat *simpler* than in our experiments. However, modeling more realistic value functions is a domain specific task as it depends on the typical queries that buyers submit. Modeling such realistic value functions is one of the directions of our future work.

4.2.1 Aggregate Demand

To illustrate combinatorial preferences of buyers, we compute aggregate demand curves that correspond to different deterministic allocations of databases. Remember that the value model we adopted for our simulations assumes, for simplicity, that marginal values v'_j and thresholds $\overline{r}_j$ of buyers depend only on the number of allocated databases but not on the identities of those databases. Thus, we can restrict our attention to considering only 10 different deterministic allocations that correspond to different numbers of allocated databases.

Figure 7 illustrates how the demand curve changes as the number of allocated databases increases (with the price on the y-axis, and the demand for rows on the x-axis). There are two effects happening in parallel as the number of allocated databases grows. First, as discussed in the experimental set-up, the threshold $\overline{r}_j$ weakly increases with the number of allocated databases (see Section 4.2). This implies that for larger number of allocated databases buyers demand weakly more rows at any price p. This shifts the demand curve to the right. Second, the marginal values of buyers also increase (see Assumption 1). As a result, for high prices, the number of buyers who can afford to buy rows at these prices increases. Consequently, the demand becomes more elastic, i.e., the slope of the demand curve gets smaller. To see this more clearly, compare the leftmost curve that corresponds to the case with a single allocated database to the second curve that corresponds to the case with two allocated databases. We see that, if the price $p > 2$, no rows are consumed in the first scenario. This happens because marginal values of all buyers are smaller than this price (remember that marginal values in this case are drawn from $U[0, 2]$). In contrast, the marginal values of the buyers in the case of two allocated databases are drawn from a distribution with a larger support, i.e., $[0, 4]$. Thus, more buyers are now reacting to price changes at the price level $p = 2$.

We see the opposite effect for low prices, i.e., the demand curve becomes less elastic for low prices (in other words, the slope of the demand curve gets larger). To understand the intuition behind this, assume, for example, that the 1st percentile of buyers' marginal values in the case of a single allocated database, is equal to p'. If we allocate an additional database, this percentile weakly increases. Consequently, in the case of two allocated databases, there are less buyers with marginal values lower than p'. This means that there are less buyers

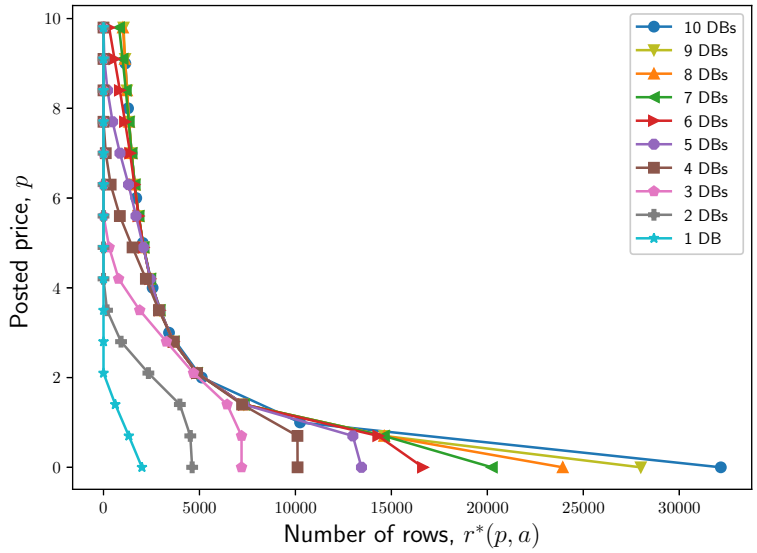

Figure 7: Aggregate demand curves in the domain with 1024 buyers and 10 databases. Different curves correspond to different numbers of allocated databases. As the number of allocated databases increases, the respective demand curve shifts to the right. The difference between two demand curves gets smaller as the number of allocated databases grows.

who switch from buying to not buying if the price changes from p' to $p' + \epsilon$, for some small $\epsilon > 0$. This results in a steeper demand curve in the area of low prices for two databases comparing to the one with a single allocated database.[21]

The main insight from this experiment is that the difference between demand curves gets smaller as the number of databases increases (see Figure 7). In practice, such "convergence" of demand curves may allow us to reduce the computational burden arising from combinatorial preferences of buyers by considering smaller domains. We conjecture that, in this case, we could bound the efficiency loss caused by approximating the demand curve of a large domain with the one corresponding to a smaller domain. However, we leave this direction to future work.

4.2.2 Expected Profit

We now study how the expected profits of sellers respond to the level of competition between sellers for producing their databases. To this end, we fix the number of databases $D = 10$ and the number of buyers $L = 1024$. We further assume that database 1 can be produced only by a single data provider s_1, who has a monopolistic ownership for the respective data. Similarly, for each database $k \leq D$, we assume that it can be produced by k different data providers who compete with each other in the BORA auction to produce the database.

21. The steep area of the demand curve in the area of low prices, however, gets "smaller" for larger number of allocated databases. This is why, in Figure 7, the steep part is hardly visible if the number of databases is larger than 5.

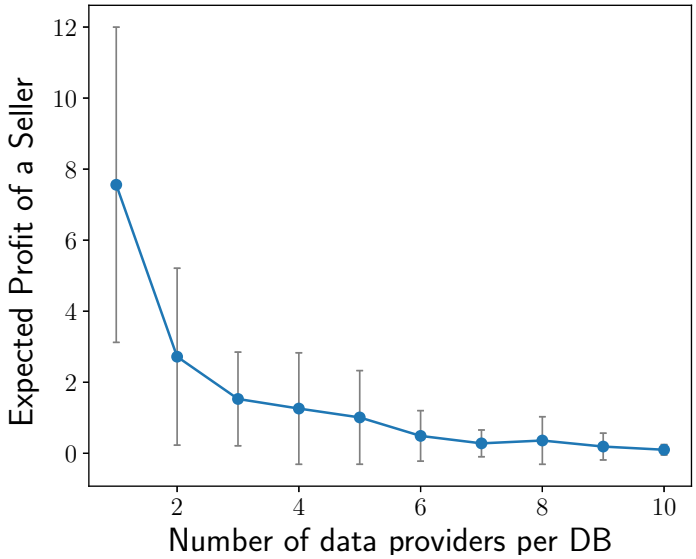

Figure 8: Expected profit of a seller in a market with 10 DBs and 1024 buyers. A seller extracts a positive expected surplus for any level of competition. The larger the number of sellers competing for producing a database, the smaller the surplus such a seller can expect to obtain.

Thus, the total number of data providers in this scenario is

$$\sum_{k=1}^{D} k = 1 + 2 + ... + 10 = 55.$$

Figure 8 demonstrates that seller s_1 enjoys the largest expected surplus. To explain this result, remember that s_1 faces no competition in the BORA auction, as she is a unique data provider for database 1. This means that the externality imposed by s_1 is potentially larger that the externality exposed by any other seller facing a stronger competition. Consequently, the payment must also be larger.

This also demonstrates that, even though that all allocated data providers can recoup their fixed costs, our market rewards data providers who *innovate* (produce original data) substantially more than those who produce databases containing *common knowledge*.

4.2.3 Shares of Social Welfare

Now, we study how the achieved social welfare is distributed between buyers and sellers. Table 2 presents the social welfare as well as the distribution of the social welfare between buyers and sellers achieved in the market as the number of buyers increases from 16 to 1024. As in Section 4.2.2, we set the number of databases $D = 10$ and we assume that database 1 can be produced by only a single data provider s_1, while database k can be produced by k different data providers.

Surplus	Number of buyers						
	16	32	64	128	256	512	1024
Buyers, %	99.5	99.7	99.8	99.9	99.9	99.9	99.9
Sellers, %	0.5	0.3	0.2	0.1	0.1	0.1	0.1
Social Welfare	8186	16585	33247	66504	133129	266344	532728

Table 2: Social welfare and the shares of buyers and sellers.

The table confirms that buyers receive the largest share of the social welfare which is expected given that the BORA auction is designed to maximize the buyers' surplus. The figure also suggests that the share of sellers' surplus stays nearly constant as the number of buyers increases. This result is expected: indeed, the payment of s_1 approaches its maximum value of $\phi^{-1}(20)$ (in this experiment, $c_1 \sim U[0, 20]$) as the number of buyers increases. At the same time, the payments of other sellers are essentially *second price* payments. These payments depend on the cost distributions F_i of sellers rather than on the number of buyers. Consequently, the total payment to be accrued to sellers does not depend much on the number of buyers. Finally, the total social welfare achieved in the market grows linearly as the number of buyers increases.

4.3 Discussion

The results of our computational experiments raise a number of interesting points regarding the design of a market for distributed data.

First, the design we proposed solves the original problem we posed. Specifically, the market brings high surplus to buyers and compensates the allocated sellers. As we showed, such a market gives stronger incentives to data providers to innovate, i.e., to produce unique data sets instead of transforming common knowledge into a structured form. The inequality in profits between unique and non-unique data providers arises from the fact that the production of a unique database reduces competition in the BORA auction. This allows data providers with monopolistic ownership of data to receive payments that are typically larger than the "second price" payment. However, even those allocated data providers who do not have access to unique data can expect positive profits.

Second, our market awards a large portion of the achieved welfare to buyers. In fact, in our experiments, even in the most extreme cases, when all data providers enjoy high profits from monopolistic ownership over their data, the buyers still get a substantial portion of the achieved welfare. In practice, providing a high share of surplus to buyers can be crucial when designing such a market. It allows for a shift in the current paradigm, meaning that the data must be free, and thus makes buyers less resistant to entering the market (some discussion on this topic in the area of Linked Data can be found, for example in Grubenmann et al. (2018b), Grubenmann et al. (2017)).

Finally, our experiments give us reason to think that the complex combinatorial preferences of buyers do not constitute an unsolvable issue. In fact, we think that the combinatorial structure can be tackled efficiently when we aggregate buyers and compute the posted price based on their aggregated preferences. In particular, we showed that, as the number of databases grows, the aggregate demand curve does not change significantly. This opens

up an opportunity to approximate the aggregate demand curve for a large domain by considering only a smaller part of it. We conjecture that this can reduce the complexity of our approach with a bounded loss in efficiency. However, we leave this direction to future work.

Limitations. In our computational experiments, we have considered buyers coming from the same value model. This means that, despite the fact that different buyers in our experiments have different marginal values v_j and thresholds $\overline{r}_j$, these values are still drawn from the same distribution. We have also assumed that all buyers are endowed with the same initial amount of money and that the preferences of all buyers depend on the number of allocated databases, not on the identities of these databases. All of these simplifying assumptions were made for better clarity of the experimental results, rather than to circumvent any complications arising in computations of the equilibrium allocation and prices. Any of these assumptions can be easily relaxed, but we expect the results to qualitatively remain the same.

5. Conclusion

In this paper, we have proposed a combinatorial market for distributed data. Our research is motivated by the increasing value of data, while the design of good mechanisms for buying and selling data has proved to be elusive. We have argued that data is different from other information goods such as music or videos because databases produced by different data providers can be joined and buyers have combinatorial values over which databases are available. The key idea behind our solution is to use two different mechanisms for the two sides of the market and to employ a fixed-point iteration algorithm for finding an outcome that balances the entire market. Our experimental results are consistent with our theoretical predictions. With a small number of buyers, it is likely that no trade happens because the buyers' values are not large enough to warrant the high fixed costs of the data providers. But, as more and more buyers arrive on the market, the probability of trade approaches one, and the posted price quickly stabilizes. We have also shown that, as the number of allocated databases increases, the marginal change in the aggregate demand gets smaller. This opens the door for future opportunities to design an approximation algorithm that would efficiently tackle the computational hardness of the equilibrium price computation procedure for larger domains. Another important discovery of our model highlights the fact that data providers who innovate by producing unique data sets can expect to receive larger rewards than those who simply provide common knowledge data. Future work can build on our model and consider various extensions, such as the dynamic arrival of sellers to the market or endogenous demand (i.e., where the number of buyers varies depending on the price). One particularly important subject of future work is the development of a realistic domain generator and large-scale simulations to study the behavior of our market mechanism under real-world conditions (e.g., with preferences of doctors or drug developers using life sciences databases (Hall et al., 2013) or of marketers in the digital marketing domain (Blue Kai, Inc., 2011)).

Appendix A. Examples

Example 3. *We use the setting of Example 2 but assume now that both sellers s_1 and s_2 are allocated, i.e., $a'' = (1,1)$. As there is now more data available to buyers, their preferences change. Assume that value functions of buyers for the new allocation are $v_1(r_1, a'') = 6\min\{r_1, 1\}$ and $v_2(r_2, a'') = \min\{r_2, 4\}$. Now, the aggregate buyer has the following aggregate value $V(r, a'')$ and demand $r^*(p, a'')$:*

$$V(r,a'') = \begin{cases} 6r, & \text{if } r \in [0,1] \\ 5+r, & \text{if } r \in [1,5] \\ 10, & \text{if } 5 \le r \end{cases} \quad r^*(p,a'') = \begin{cases} 5, & \text{if } p \in (0,1) \\ [1,5], & \text{if } p = 1 \\ 1, & \text{if } p \in (1,6) \\ [0,1], & \text{if } p = 6 \\ 0, & \text{if } 6 < p. \end{cases}$$

In this case, the positive externality imposed by s_2 is $ext_2(a'', p) = V(r^*(p, a''), a'') - V(r^*(p, a), a)$. Again, if $p = 1 - \epsilon$, then the market demand $r^*(p, a'') = 5$ and the aggregate value of having both databases is $V(r^*(p, a''), a'') = V(5, a'') = 10$. Then, the externality imposed by s_2 is $ext_2(a'', p) = V(r^*(p, a''), a'') - V(r^*(p, a), a) = 10 - 6 = 4$. If we now assume that for allocation $a''' = (0,1)$ agents have same preferences as for allocation $a = (1,0)$, then $ext_1(a'', p) = V(r^*(p, a''), a'') - V(r^*(p, a'''), a''') = 10 - 6 = 4$. Thus, the induced value of the databases are $W_1(p, a'') = \frac{4 \cdot 10}{8} = 5$, $W_2(p, a'') = \frac{4 \cdot 10}{8} = 5$. Observe, that the presence of s_2 increased the induced value $W_1(p, a'')$ for the database of s_1.

Example 4. *Consider a domain with a single buyer, $L = 1$. Assume that there are $N = 2$ sellers each producing a single database, i.e., $D = 2$. Let $c_1, c_2 \sim U[0, 2]$ and $c_1 = 1.5$, $c_2 = 0.5$. In this case, the virtual cost function for both sellers is $\phi(c) = c + \frac{F(c)}{f(c)} = 2c$; consequently, $\phi_1(c_1) = 3$ and $\phi_2(c_2) = 1$. Assume that the value function of the buyer is $v_1(r_1, a) = 5\min\{r_1, 1\}$ if both databases are allocated (i.e., $a = (1,1)$) and $v_1(r_1, a) = 0$ otherwise. The buyer has an endowment $e = 4$. As there is only a single buyer, the aggregate value function corresponds to the value function of this buyer, i.e., $V(r, a) = v_1(r, a)$. The endowment of the aggregate buyer is $E = e$.*

Let us now compute the induced values of both databases. First, $ext_1(a, p) = ext_2(a, p) = 5$ for all $p \le 5$ and for $a = (1,1)$. Also $ext_1(a, p) = ext_2(a, p) = 0$ if $a \neq (1,1)$ or if $p > 5$. Thus, $W_1(a, p) = W_2(a, p) = \frac{1}{2} \cdot 5 = 2.5$ for any $p \le 5$ if $a = (1,1)$ and $W_1(a, p) = W_2(a, p) = 0$ for other cases. Obviously, the solution to the allocation problem is $a^* = (a_1^*, a_2^*) = (1,1)$. In this case, the objective is $2.5 + 2.5 - 3 - 1 = 1$ for any $p \le 5$. The payments are computed as follows: $t_1 = \frac{1}{2}(3 + 1 - 0) = 2$, $t_2 = \frac{1}{2}(1 + 1 - 0) = 1$.

Observe that if we run instead two BORA auctions for each of the databases separately, we first would not allocate the first database as its virtual cost $\phi_1(c_1) = 3$ is larger than the induced value of the database $W_1(a, p) = 2.5$. Consequently, the second database would also not be allocated as the buyer has a positive value only for both databases.

Finally, if we set $p = 3$, then the buyer would decide to pay $e - m_1^*(p, a) = 3$ for a single row of answers for his query, $r_1^*(p, a) = 1$. Such a price makes the overall market balanced as the total payment to the sellers must be $t_1 + t_2 = 3$.

Example 5. *In this example, we would like to demonstrate that even if there exist multiple non-trivial equilibria, our mechanism finds the "best" one, i.e., the equilibrium with the largest surplus for buyers.*

Consider a domain with a single buyer, $L = 1$*. Assume that there are* $N = 2$ *sellers each producing a single database. Let* $c_1, c_2 \sim U[0,1]$ *and* $c_1 = c_2 = 0.5$*. In this case, the virtual cost functions are* $\phi_1(c_1) = \phi_2(c_2) = 1$*. Assume that the value function of the buyer is* $v_1(r_1, a) = 2\min\{r_1, 6\}$ *if both databases are allocated (i.e.,* $a = (1,1)$*) and* $v_1(r_1, a) = 2\min\{r_1, 2\}$ *if only one database is allocated. The buyer has an endowment* $e = 10$*. The aggregate value function is* $V(r, a) = v_1(r, a)$ *(the endowment of the aggregate buyer is* $E = e$*).*

Suppose Algorithm 1 starts with $p = 0$*. In this case, both databases must be allocated. However, the market is not budget balanced as the buyer pays 0. Assume now that after several iterations of Algorithm 1, the price increases to* $p = 1$*. In this case, an allocation* $a = (1,0)$ *makes the market budget balanced. Indeed, in this case,* $W_1(a, p) = 4$ *while* $W_2(a, p) = 0$*. Consequently, the virtual surplus is* $\tilde{S} = 4 - 1 = 3$ *and the payments are* $t_1 = 2$ *and* $t_2 = 0$*. Given this price, the buyer decides to buy two rows and thus pays the total amount of* 2 *which implies budget balancedness. In this equilibrium, the buyer gets a surplus of 2.*

Observe however, that the allocation $a = (1,0)$ *does not maximize the virtual surplus given the price* $p = 1$*. Instead, the BORA auction would allocate both databases. This would lead to a virtual surplus of* $\tilde{S} = 10$ *and payments* $t_1 = 4$*,* $t_2 = 4$*. This, makes the market unbalanced as buyers can pay only 6. Consequently, the price must increase up to* $p = 4/3$ *for the market to become budget balanced. In the new equilibrium, the buyer gets a surplus of 4 and pays the total amount of* $t_1 + t_2 = 8$ *to sellers.*

Appendix B. Proofs

Let $G_i(g, c_i) = \int_{c_{-i}} g_i(c_i, c_{-i}) f_{-i}(c_{-i}) dc_{-i}$ denote the ex-interim allocation of s_i.

Definition 7. *A mechanism $\mathcal{A} = \langle g, h \rangle$ is **feasible** if it satisfies BNIC, IR, and $\sum_{i \in i(k)} g_i(c) \leq 1$, $\forall k = 1, ..., D$, $g_i(c) \geq 0$, $\forall i = 1, ..., N$.*

Lemma B.1. *A mechanism $\mathcal{A} = \langle g, h \rangle$ is feasible if and only if the following conditions hold:*

1. *$c_i \leq q_i$ implies $G_i(g, q_i) \leq G_i(g, c_i)$ for any q_i, $\forall c_i \in [\alpha_i, \beta_i]$, $i = 1, ..., N$,*
2. *$\mathbb{E}_{f_{-i}}[u_i(g, c_i, h)] = \mathbb{E}_{f_{-i}}[u_i(g, \beta_i, h)] + \int_{c_i}^{\beta_i} G_i(g, q_i) dq_i$,*
3. *$\mathbb{E}_{f_{-i}}[u_i(g, \beta_i, h)] \geq 0$ for all $i = 1, ..., N$,*

and $\sum_{i \in i(k)} g_i(c) \leq 1$, $g_i(c) \geq 0$ for all $k = 1, ..., D$, $\forall c \in \prod_{i=1}^{N} [\alpha_i, \beta_i]$.

Proof. The proof repeats the respective proof provided in (Myerson, 1981) for a reverse auction setting.

□

In words, the first condition of the previous Lemma means *monotonicity of ex-interim allocation* while the second and the third conditions are more technical and will be used in derivation of the surplus optimal mechanism.

Now, remember that the probabilistic allocation of sellers $g(c)$ induces a probabilistic allocation of databases. We can define the expected induced value of a database k as follows:

Definition 8. *The **expected induced value** of the database k given the probabilistic allocation $g(c)$ of sellers is*

$$\mathbb{E}_{g(c)}\Big[W_k(a, p)\Big] = \sum_{a \in \{0,1\}^N} \prod_{i=1}^{N} g_i^{a_i}(c)(1 - g_i(c))^{1 - a_i} W_k(a, p).$$

In a setting when multiple sellers compete for producing a database ℓ, an assignment of the full allocation probability $\gamma \geq 0$ to only one of them leads to weakly higher expected induced values of all databases than any other assignment of γ. The following lemma shows this fact formally:

Lemma B.2. *Let $g(c)$ be a probabilistic allocation of sellers such that $\sum_{i \in i(\ell)} g_i(c) = \gamma$ for some $\ell \leq D$, $0 \leq \gamma \leq |i(\ell)|$. Then, for any allocation $g'(c)$ such that $g'_q = \min\{1, \gamma\}$ for some $q \in i(\ell)$, $g'_s = 0$ $\forall s \in i(\ell) \backslash q$, and $g'_j = g_j$ $\forall j \notin i(\ell)$ we have*

$$\mathbb{E}_{g'(c)}\Big[W_k(a, p)\Big] \geq \mathbb{E}_{g(c)}\Big[W_k(a, p)\Big], \quad \forall k \leq D.$$

Proof. W. l. o. g. let $s_1, ..., s_d$ be sellers producing the database ℓ (here, $d = |i(\ell)|$). We first introduce some helpful notation. Specifically, let $a_{1:d} = (a_1, ..., a_d) \in \{0, 1\}^d$ and $a_{-1:d} = (a_{d+1}, ..., a_N) \in \{0, 1\}^{N-d}$ be the allocation of the first d sellers and of the rest of the sellers respectively. Thus, we can rewrite $a = (a_{1:d}, a_{-1:d})$.

In this case, from Lemma B.3 it follows that $\forall a_{-1:d}$, $\forall p$, for any two $a_{1:d}, a'_{1:d} : ||a_{1:d}|| > 0, ||a'_{1:d}|| > 0$, we have $W_k((a_{1:d}, a_{-1:d}), p) = W_k((a'_{1:d}, a_{-1:d}), p)$ for any $k \leq D$.

Consider now the expected induced value of a database k under the probabilistic allocation $g(c)$:

$$\mathbb{E}_{g(c)}\Big[W_k(a,p)\Big] = \sum_{a\in\{0,1\}^N} \prod_{i=1}^{N} g_i^{a_i}(c)(1-g_i(c))^{1-a_i} W_k(a,p) =$$

$$\sum_{a_{-1:d}\in\{0,1\}^{N-d}} (1-g_1)\cdot ...\cdot (1-g_d) \prod_{i=1}^{N-d} g_{d+i}^{a_{-1:d}(i)}(1-g_{d+i})^{1-a_{-1:d}(i)} W_k((0,...,0,a_{-1:d}),p) +$$

$$\sum_{a_{-1:d}\in\{0,1\}^{N-d}} g_1\cdot ...\cdot (1-g_d) \prod_{i=1}^{N-d} g_{d+i}^{a_{-1:d}(i)}(1-g_{d+i})^{1-a_{-1:d}(i)} W_k((1,...,0,a_{-1:d}),p) +$$

$$...$$

$$\sum_{a_{-1:d}\in\{0,1\}^{N-d}} g_1\cdot ...\cdot g_d \prod_{i=1}^{N-d} g_{d+i}^{a_{-1:d}(i)}(1-g_{d+i})^{1-a_{-1:d}(i)} W_k((1,...,1,a_{-1:d}),p) =$$

$$\sum_{a_{-1:d}\in\{0,1\}^{N-d}} (1-g_1)\cdot ...\cdot (1-g_d) \prod_{i=1}^{N-d} g_{d+i}^{a_{-1:d}(i)}(1-g_{d+i})^{1-a_{-1:d}(i)} W_k((0,...,0,a_{-1:d}),p) +$$

$$\sum_{a_{-1:d}\in\{0,1\}^{N-d}} (1-(1-g_1)\cdot ...\cdot (1-g_d)) \prod_{i=1}^{N-d} g_{d+i}^{a_{-1:d}(i)}(1-g_{d+i})^{1-a_{-1:d}(i)} W_k((a_{1:d},a_{-1:d}),p).$$

Here, $||a_{1:d}|| > 0$ implies that there exists $q \leq d$, s.t., $a_{1:d}(q) = 1$. We now can rewrite the expression above as

$$\sum_{a_{-1:d}\in\{0,1\}^{N-d}} (1-g_1)\cdot ...\cdot (1-g_d) \prod_{i=1}^{N-d} g_{d+i}^{a_{-1:d}(i)}(1-g_{d+i})^{1-a_{-1:d}(i)} W_k((0,...,0,a_{-1:d}),p) +$$

$$\sum_{a_{-1:d}\in\{0,1\}^{N-d}} \prod_{i=1}^{N-d} g_{d+i}^{a_{-1:d}(i)}(1-g_{d+i})^{1-a_{-1:d}(i)} W_k((a_{1:d},a_{-1:d}),p) -$$

$$\sum_{a_{-1:d}\in\{0,1\}^{N-d}} (1-g_1)\cdot ...\cdot (1-g_d) \prod_{i=1}^{N-d} g_{d+i}^{a_{-1:d}(i)}(1-g_{d+i})^{1-a_{-1:d}(i)} W_k((a_{1:d},a_{-1:d}),p) =$$

$$\sum_{a_{-1:d}\in\{0,1\}^{N-d}} \prod_{i=1}^{N-d} g_{d+i}^{a_{-1:d}(i)}(1-g_{d+i})^{1-a_{-1:d}(i)} W_k((a_{1:d},a_{-1:d}),p) +$$

$$\sum_{a_{-1:d}\in\{0,1\}^{N-d}} (1-g_1)\cdot ...\cdot (1-g_d) \prod_{i=1}^{N-d} g_{d+i}^{a_{-1:d}(i)}(1-g_{d+i})^{1-a_{-1:d}(i)} \times$$

$$\underbrace{(W_k((0,...,0,a_{-1:d}),p) - W_k((a_{1:d},a_{-1:d}),p))}_{\leq 0}.$$

Notice that the last term in the expression above is non-positive due to Assumption 3. At the same time, the first summand does not depend on $g_1, ..., g_d$. Thus, the problem now is to find such an assignment of $g_1, ..., g_d$ that is feasible (i.e., $0 \leq g_i \leq 1$, $\forall i \in i(\ell)$ and $\sum_{i \in i(\ell)} g_i = \gamma$) and that minimizes $(1 - g_1) \cdot ... \cdot (1 - g_d)$. We claim that $g_q = \min\{\gamma, 1\}$ for some $q \in i(\ell)$ and $g_s = 0$ $\forall s \in i(\ell) \backslash q$ is such an assignment.

To see this, let's consider only the case when $q = 1$ (all other cases are symmetric). We proceed by induction in d. If $d = 1$, the statement is trivial. Indeed, in this case, $g_1 = \min\{1, \gamma\}$ minimizes $(1 - g_1)$. Now consider the case $d = 2$. In this case, we are solving the following problem:

$$\begin{aligned} & \min_{g_1, g_2} (1 - g_1)(1 - g_2) \\ \texttt{s.t.} \quad & g_1 + g_2 = \gamma \\ & g_1, g_2 \in [0, 1]. \end{aligned}$$

If $\gamma \geq 1$, let us rewrite the objective function as $\min_{g_2}(1-\gamma+g_2)(1-g_2) = \min_{g_2} 1-g_2^2-\gamma+\gamma g_2$. In this case, the concave objective function is minimized at the boundary of the $[0, 1]$ interval, namely when $g_2 = 0$ (respectively, $g_1 = \min\{1, \gamma\} = 1$). The optimal objective value in this case is 0.

If $\gamma < 1$, let us rewrite the objective as $\min_{g_1}(1-\gamma+g_1)(1-g_1) = \min_{g_1} 1-g_1^2-\gamma+\gamma g_1$. In this case, $g_1 = \gamma$ minimizes the objective function. The optimal objective value in this case is $1 - \gamma$.

Assume now that the statement is true for some r, $1 \leq r \leq d$, i.e., $(1 - g_1) \cdot ... \cdot (1 - g_r)$ is minimized by setting $g_1 = \min\{1, \gamma\}$. Consider the following problem for $r + 1$:

$$\begin{aligned} & \min_{g_1, ..., g_{r+1}} (1 - g_1) \cdot ... \cdot (1 - g_{r+1}) \\ \texttt{s.t.} \quad & \sum_{i=1}^{r+1} g_i = \gamma \\ & g_i \in [0, 1] \ \ \forall i = 1, ..., r + 1. \end{aligned}$$

We can rewrite it as

$$\begin{aligned} & \min_{g_1, ..., g_{r+1}} (1 - g_1)\Big((1 - g_2)... \cdot (1 - g_{r+1})\Big) \\ \texttt{s.t.} \quad & \sum_{i=2}^{r+1} g_i = \gamma - g_1 \\ & g_i \in [0, 1] \ \ \forall i = 1, ..., r + 1. \end{aligned}$$

From the induction hypothesis, it follows that for any $0 \leq g_1 \leq \gamma$ setting $g_2 = \gamma - g_1$, $g_3 = ... = g_{r+1} = 0$ minimizes $(1 - g_2) \cdot ... \cdot (1 - g_{r+1})$. Therefore, the problem is to find such g_1, g_2 that solve

$$\begin{aligned} & \min_{g_1, g_2} (1 - g_1)(1 - g_2) \\ \texttt{s.t.} \quad & g_2 = \gamma - g_1 \\ & g_1, g_2 \in [0, 1]. \end{aligned}$$

As we have shown above, the solution to this problem is $g_1 = \min\{1, \gamma\}$, $g_2 = 0$. Q.E.D.

□

Lemma 3.4. *Let* $g : \mathbb{R}^N_{\geq 0} \to [0,1]^N$ *maximizes*

$$\mathbb{E}_f\Big[\sum_{k=1}^{D}\Big(\mathbb{E}_{g(c)}\Big[W_k(a,p)\Big] - \sum_{i\in i(k)}\Big(c_i + \frac{F_i(c_i)}{f_i(c_i)}\Big)g_i(c)\Big)\Big]$$

subject to monotonicity of ex-interim allocation and constraints (13), (14). *Let also* $h_i(c) = g_i(c)c_i + \int_{c_i}^{\beta_i} g_i(q_i, c_{-i})dq_i$ *for every* $i = 1, ..., N$, $\forall c$. *Then* $\mathcal{A} = \langle g, h\rangle$ *is an optimal surplus maximizing reverse auction.*

Proof. The proof is similar to the one presented in (Myerson, 1981). Consider the surplus of the auctioneer.

$$\mathbb{E}_f[U(g,h)] = \mathbb{E}_f\Big[\sum_{k=1}^{D}\mathbb{E}_{g(c)}\Big[W_k(a,p)\Big] - \sum_{i=1}^{N}h_i(c)\Big] = \int_c\Big(\sum_{k=1}^{D}\mathbb{E}_{g(c)}\Big[W_k(a,p)\Big] -$$

$$\sum_{i=1}^{N}h_i(c)\Big)f(c)dc = \int_c\sum_{k=1}^{D}\mathbb{E}_{g(c)}\Big[W_k(a,p)\Big]f(c)dc - \int_c\sum_{i=1}^{N}h_i(c)f(c)dc =$$

$$\sum_{k=1}^{D}\int_c\mathbb{E}_{g(c)}\Big[W_k(a,p)\Big]f(c)dc - \sum_{i=1}^{N}\int_c c_ig_i(c)f(c)dc + \sum_{i=1}^{N}\int_c c_ig_i(c)f(c)dc - \int_c\sum_{i=1}^{N}h_i(c)f(c)dc =$$

$$\sum_{k=1}^{D}\int_c\mathbb{E}_{g(c)}\Big[W_k(a,p)\Big]f(c)dc - \sum_{k=1}^{D}\sum_{i\in i(k)}\int_c c_ig_i(c)f(c)dc + \sum_{i=1}^{N}\int_c c_ig_i(c)f(c)dc - \sum_{i=1}^{N}\int_c h_i(c)f(c)dc =$$

$$\underbrace{\sum_{k=1}^{D}\int_c\Big(\mathbb{E}_{g(c)}\Big[W_k(a,p)\Big] - \sum_{i\in i(k)}c_ig_i(c)\Big)f(c)dc}_{A} + \sum_{i=1}^{N}\int_c\Big(c_ig_i(c) - h_i(c)\Big)f(c)dc =$$

$$A + \sum_{i=1}^{N}\int_{c_i}\int_{c_{-i}}(c_ig_i(c) - h_i(c))f_i(c_i)f_{-i}(c_{-i})dc_idc_{-i} = A - \sum_{i=1}^{N}\int_{c_i}\mathbb{E}_{f_{-i}}[u_i(g,c_i,h)]f_i(c_i)dc_i.$$

Given that we are looking for a feasible mechanism, we can rewrite $\mathbb{E}_{f_{-i}}[u_i(g, c_i, h)] = \mathbb{E}_{f_{-i}}[u_i(g, \beta_i, h)] + \int_{c_i}^{\beta_i} G_i(g, q_i)dq_i$ (see condition 2 of Lemma B.1). Therefore,

$$\mathbb{E}_f[U(g,h)] = A - \sum_{i=1}^{N} \int_{c_i} \mathbb{E}_{f_{-i}}[u_i(g,\beta_i,h)]f_i(c_i)dc_i - \sum_{i=1}^{N} \int_{c_i}\int_{c_i}^{\beta_i} G_i(g,q_i)dq_i f_i(c_i)dc_i =$$

$$A - \underbrace{\sum_{i=1}^{N} \mathbb{E}_{f_{-i}}[u_i(g,\beta_i,h)]}_{B} - \sum_{i=1}^{N} \int_{c_i}\int_{c_i}^{\beta_i} G_i(g,q_i)dq_i f_i(c_i)dc_i = A - B - \sum_{i=1}^{N} \int_{c_i}\int_{c_i}^{\beta_i} G_i(g,q_i)dq_i dF_i(c_i) =$$

$$A - B - \sum_{i=1}^{N} \Big[F_i(c_i)\int_{c_i}^{\beta_i} G_i(g,q_i)dq_i\Big|_{\alpha_i}^{\beta_i} - \int_{\alpha_i}^{\beta_i} F_i(c_i)d\int_{c_i}^{\beta_i} G_i(g,q_i)dq_i\Big] =$$

$$A - B - \sum_{i=1}^{N} \Big[0 + \int_{\alpha_i}^{\beta_i} F_i(c_i)G_i(g,c_i)dc_i\Big] = A - B - \sum_{i=1}^{N} \int_{\alpha_i}^{\beta_i}\int_{c_{-i}} g_i(c_i,c_{-i})f_{-i}(c_{-i})dc_{-i}F_i(c_i)dc_i =$$

$$A - B - \sum_{i=1}^{N} \int_{c} g_i(c_i,c_{-i})\frac{F_i(c_i)}{f_i(c_i)}f(c)dc = \sum_{k=1}^{D}\int_{c}\Big(\mathbb{E}_{g(c)}\Big[W_k(a,p)\Big] - \sum_{i\in i(k)}\Big(c_i + \frac{F_i(c_i)}{f_i(c_i)}\Big)g_i(c)\Big)f(c)dc - B.$$

From Lemma B.1 it follows that

$$B = \sum_{i=1}^{N} \mathbb{E}_{f_{-i}}[u_i(g,\beta_i,h)] = \sum_{i=1}^{N}\Big(\mathbb{E}_{f_{-i}}[u_i(g,c_i,h)] - \int_{c_i}^{\beta_i} G_i(g,q_i)dq_i\Big) =$$

$$\sum_{i=1}^{N}\Big(\int_{c_{-i}}(h_i(c) - g_i(c)c_i)f_{-i}(c_{-i})dc_{-i} - \int_{c_i}^{\beta_i}\int_{c_{-i}} g_i(q_i,c_{-i})f_{-i}(c_{-i})dc_{-i}dq_i\Big) =$$

$$\sum_{i=1}^{N}\Big(\int_{c_{-i}}(h_i(c) - g_i(c)c_i - \int_{c_i}^{\beta_i} g_i(q_i,c_{-i})dq_i)f_{-i}(c_{-i})dc_{-i}\Big).$$

From the third condition of Lemma B.1 it follows that $h_i(c) - g_i(c)c_i - \int_{c_i}^{\beta_i} g_i(q_i, c_{-i})dq_i \geq 0$. This means that the payment rule that maximizes the expected surplus of the auctioneer, $\mathbb{E}_f[U(g,h)]$, must satisfy $h_i(c) - g_i(c)c_i - \int_{c_i}^{\beta_i} g_i(q_i,c_{-i})dq_i = 0$. Consequently, $h_i(c) = g_i(c)c_i + \int_{c_i}^{\beta_i} g_i(q_i,c_{-i})dq_i$.

This means that the problem now is reduced to finding such an allocation function g that maximizes

$$\mathbb{E}_f\Big[\sum_{k=1}^{D}\Big(\mathbb{E}_{g(c)}\Big[W_k(a,p)\Big] - \sum_{i\in i(k)}\Big(c_i + \frac{F_i(c_i)}{f_i(c_i)}\Big)g_i(c)\Big)\Big].$$

Q.E.D. □

Lemma 3.5. *Let $g^*(c)$ be a solution of (16). Then, if for a database ℓ there exists a seller s_i with $i \in i(\ell)$ such that $g_i^*(c) > 0$, then $\sum_{i\in i(\ell)} g_i^*(c) = 1$.*

Proof. Assume that there are d sellers $s_1, ..., s_d$ who produce the database ℓ, i.e. $1, ..., d \in i(\ell)$. Assume that $g_1^*(c) > 0$ but $\sum_{i\in i(\ell)} g_i^*(c) < 1$. Consequently, $g_1^*(c) < 1$. Consider the first term under the expectation in Equation 16:

$$\sum_{k=1}^{D} \mathbb{E}_{g(c)}\Big[W_k(a,p)\Big] = \sum_{k=1}^{D} \sum_{a_{-1:d}\in\{0,1\}^{N-d}} g_1 \cdot ... \cdot g_d \prod_{i=1}^{N-d} g_i^{a_{-1:d}(i)}(1-g_i)^{1-a_i} W_k((1,...,1,a_{-1:d}),p) +$$
$$\sum_{k=1}^{D} \sum_{a_{-1:d}\in\{0,1\}^{N-d}} g_1 \cdot ... \cdot (1-g_d) \prod_{i=1}^{N-d} g_i^{a_{-1:d}(i)}(1-g_i)^{1-a_i} W_k((1,...,0,a_{-1:d}),p) +$$
$$...$$
$$\sum_{k=1}^{D} \sum_{a_{-1:d}\in\{0,1\}^{N-d}} (1-g_1) \cdot ... \cdot (1-g_d) \prod_{i=1}^{N-d} g_i^{a_{-1:d}(i)}(1-g_i)^{1-a_i} W_k((0,...,0,a_{-1:d}),p).$$

Observe, that the sum above is linear in g_1, i.e., we can rewrite it as

$$\sum_{k=1}^{D} \mathbb{E}_{g(c)}\Big[W_k(a,p)\Big] = g_1\lambda(g_{-1}) + \gamma(g_{-1}),$$

where $\lambda(g_{-1})$ and $\gamma(g_{-1})$ are independent of g_1. Now, we can rewrite Equation 16 as follows

$$\mathbb{E}_f\Big[\sum_{k=1}^{D}\Big(\mathbb{E}_{g^*(c)}\Big[W_k(a,p)\Big] - \sum_{i\in i(k)}\Big(c_i + \frac{F_i(c_i)}{f_i(c_i)}\Big)g_i^*(c)\Big)\Big] =$$

$$\mathbb{E}_f\Big[g_1^*(c)\lambda(g_{-1}^*) - \phi_1(c_1)g_1^*(c) - ... - \phi_d(c_d)g_d^*(c) - \sum_{k=2}^{D}\sum_{i\in i(k)}\phi_i(c_i)g_i^*(c) + \gamma(g_{-1}^*)\Big] =$$

$$\mathbb{E}_f\Big[g_1^*(c)\Big(\lambda(g_{-1}^*) - \phi_1(c_1)\Big) \underbrace{- \phi_2(c_2)g_2^*(c) ... - \phi_d(c_d)g_d^*(c) - \sum_{k=2}^{D}\sum_{i\in i(k)}\phi_i(c_i)g_i^*(c) + \gamma(g_{-1}^*)}_{\texttt{does not depend on } g_1^*(c)}\Big].$$

Here, $g_1^*(c) > 0$ implies that $\lambda(g_{-1}^*) \geq \phi_1(c_1)$. Thus, if for some positive $\epsilon \leq 1 - g_1^*$ we take any $g'(c)$ such that $g_1'(c) = g_1^*(c) + \epsilon$ and $g_i'(c) = g_i^*(c)$ for all $i \neq 1$, then

$$\mathbb{E}_f\Big[\sum_{k=1}^{D}\Big(\mathbb{E}_{g'(c)}\Big[W_k(a,p)\Big] - \sum_{i\in i(k)}\Big(c_i + \frac{F_i(c_i)}{f_i(c_i)}\Big)g_i'(c)\Big] =$$

$$\mathbb{E}_f\Big[(g_1^*(c)+\epsilon)\Big(\lambda(g_{-1}^*) - \phi_1(c_1)\Big) - ... - \phi_d(c_d)g_d^*(c) - \sum_{k=2}^{D}\sum_{i\in i(k)}\phi_i(c_i)g_i^*(c) + \gamma(g_{-1}^*)\Big] \geq$$

$$\mathbb{E}_f\Big[g_1^*(c)\Big(\lambda(g_{-1}^*) - \phi_1(c_1)\Big) - ... - \phi_d(c_d)g_d^*(c) - \sum_{k=2}^{D}\sum_{i\in i(k)}\phi_i(c_i)g_i^*(c) + \gamma(g_{-1}^*)\Big].$$

This means that $g^*(c)$ cannot be an optimal solution. Contradiction. Q.E.D. □

Lemma B.3. *For any database k and any price p, for any $a, a' \in \{0,1\}^N$ such that $\forall \ell \notin i(k)$: $a_\ell = a'_\ell$ and $\exists s, q \in i(k) : a_q = 1, a'_s = 1$ it follows*

$$W_k(a,p) = W_k(a',p).$$

Proof. From our assumption that buyers are indifferent about the identities of sellers it follows that $\forall k$, $\forall b_j$, $\forall p$ and for any a and a' satisfying the conditions above, we have $r_j^*(p,a) = r_j^*(p,a')$ and $m_j^*(p,a) = m_j^*(p,a')$. Consequently, $r^*(p,a) = \sum_{j=1}^{L} r_j^*(p,a) = \sum_{j=1}^{L} r_j^*(p,a') = r^*(p,a')$; similarly, $m^*(p,a) = m^*(p,a')$.

It follows that $\pi(z,a) = \pi(z,a')$. This results in $V(r,a) = V(r,a')$ and consequently in $ext_k(a,p) = ext_k(a',p)$. From this it immediately follows that $W_k(a,p) = W_k(a',p)$. Q.E.D. □

Proposition 3.6. *The share of the buyers' surplus achieved in Algorithm 1 auction is lower bounded by zero.*

Proof. We prove this statement by providing an example of the domain in which buyers reach zero surplus. Consider a domain with a single buyer, $L = 1$. Assume that there are $N = 2$ sellers each producing a single database, i.e., $D = 2$. Let $c_1, c_2 \sim U[0,3]$ and $c_1 = c_2 = 0$. The virtual cost function for both sellers is $\phi(c) = c + \frac{F(c)}{f(c)} = 2c$; thus, $\phi_1(c_1) = \phi_2(c_2) = 0$.

Assume that the value function of the buyer is $v_1(r_1, a) = 5 \min\{r_1, 1\}$ if both databases are allocated (i.e., $a = (1,1)$) and $v_1(r_1, a) = 0$ otherwise. The buyer's endowment is $e = 5$. With a single buyer, the aggregate value function $V(r,a) = v_1(r,a)$ and the aggregate endowment $E = e$.

Let us now compute the induced values. First, $ext_1(a,p) = ext_2(a,p) = 5$ for all $p \leq 5$ and $a = (1,1)$. Also $ext_1(a,p) = ext_2(a,p) = 0$ if $a \neq (1,1)$ or if $p > 5$. Thus, $W_1(a,p) = W_2(a,p) = 2.5$ for any $p \leq 5$ if $a = (1,1)$ and $W_1(a,p) = W_2(a,p) = 0$ otherwise.

Given price p, the allocation problem in this case is $\max_{a_1,a_2} \Big\{ W_1((a_1,a_2),p) + W_2((a_1,a_2),p) - \phi_1(c_1)a_1 - \phi_2(c_2)a_2 \Big\}$. The solution to this problem is $(a_1^*, a_2^*) = (1,1)$. In this case, the objective value is $2.5 + 2.5 - 0 \cdot 1 - 0 \cdot 1 = 5$ for any $p \leq 5$. Payments are computed as follows: $t_1 = \frac{1}{2}(0 + 5 - 0) = 2.5$, $t_2 = \frac{1}{2}(0 + 5 - 0) = 2.5$. Setting the price $p = 5$, the market becomes balanced. In this case, the buyer pays $5 = (t_1 + t_2)$ for a single row of answers for his query, $r_1^*(p,a) = 1$. However, the buyer's surplus is 0. Q.E.D. □

References

Agarwal, A., Dahleh, M., & Sarkar, T. (2019). A marketplace for data: An algorithmic solution. Tech. rep., Massachusetts Institute of Technology, Working Paper.

Bakos, Y., & Brynjolfsson, E. (1999). Bundling information goods: Pricing, profits, and efficiency. *Management Science*, *45*(12), 1613--1630.

Balazinska, M., Howe, B., Koutris, P., Suciu, D., & Upadhyaya, P. (2013). A discussion on pricing relational data. *In Search of Elegance in the Theory and Practice of Computation*, *8000*, 167--173.

Bernstein, A., Hendler, J., & Noy, N. (2016). A new look at the semantic web. *Commun. ACM*, *59*(9), 35--37.

Blue Kai, Inc. (2011). Whitepaper: Data management platforms demystified. Tech. rep..

Buil-Aranda, C., Hogan, A., Umbrich, J., & Vandenbussche, P.-Y. (2013). Sparql web-querying infrastructure: Ready for action?. In *Proceedings of the 12th International Semantic Web Conference - Part II*, ISWC '13, pp. 277--293, New York, NY, USA. Springer-Verlag New York, Inc.

Chambers, C. P., & Echenique, F. (2009). Supermodularity and preferences. *Journal of Economic Theory*, *144*(3), 1004 -- 1014.

Cheng, J. Q., & Wellman, M. P. (1998). The walras algorithm: A convergent distributed implementation of general equilibrium outcomes. *Comput. Econ.*, *12*(1), 1--24.

Deep, S., & Koutris, P. (2016). The design of arbitrage-free data pricing schemes. Tech. rep., University of Wisconsin-Madison, Working Paper.

Ferrucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A. A., Lally, A., Murdock, J. W., Nyberg, E., Prager, J., Schlaefer, N., & Welty, C. (2010). Building Watson: An overview of the DeepQA project.. *AI Magazine*, *31(3)*, 59--79.

Goldberg, A., & Hartline, J. (2001). Competitive auctions for multiple digital goods. *Springer Berlin Heidelberg, Berlin, Heidelberg*, 416--427.

Goldberg, A. V., & Hartline, J. D. (2003). Envy-free auctions for digital goods. In *Proceedings of the 4th ACM Conference on Electronic Commerce*, EC '03, pp. 29--35, New York, NY, USA. ACM.

Goldberg, A. V., Hartline, J. D., & Wright, A. (2001). Competitive auctions and digital goods. In *Proceedings of the Twelfth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '01, pp. 735--744, Philadelphia, PA, USA. Society for Industrial and Applied Mathematics.

Greg, M. (2011). Inside the bloomberg machine. WallStreet and Technology, http://www.wallstreetandtech.com/trading-technology/inside-the-bloomberg-machine/d/d-id/1264634?

Grubenmann, T., Bernstein, A., Moor, D., & Seuken, S. (2018a). Fedmark: A marketplace for federated data on the web. Tech. rep., University of Zurich, Working Paper.

Grubenmann, T., Bernstein, A., Moor, D., & Seuken, S. (2018b). Financing the web of data with delayed-answer auctions. In *Proceedings of the 2018 World Wide Web*

Conference, WWW '18, pp. 1033--1042, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Grubenmann, T., Dell'Aglio, D., Bernstein, A., Moor, D., & Seuken, S. (2017). Decentralizing the Semantic Web: who will pay to realize it?. In *Proceedings of the Workshop on Decentralizing the Semantic Web (DeSemWeb)*.

Hall, A. S., Shan, Y., Lushington, G., & Visvanathan, M. (2013). An overview of computational life science databases and exchange formats of relevance to chemical biology research. *Combinatorial Chemistry and High Throughput Screening*, *16*(3), 189--198.

HCLS (2001). Semantic Web Health Care and Life Sciences (HCLS) Interest Group. https://www.w3.org/2001/sw/hcls/.

Koutris, P., Upadhyaya, P., Balazinska, M., Howe, B., & Suciu, D. (2013). Toward practical query pricing with querymarket. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*, pp. 613--624.

Koutris, P., Upadhyaya, P., Balazinska, M., Howe, B., & Suciu, D. (2015). Query-based data pricing. In *Journal of the ACM (JACM)*, Vol. 62.

Mas-Colell, A., Whinston, M. D., & Green, J. R. (1995). *Microeconomic Theory*. Oxford University Press, New York.

McKinsey report (2016). Creating a successful Internet of Things Data Marketplace. http://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/creating-a-successful-internet-of-things-data-marketplace?cid=soc-web.

Microsoft Azure Data Marketplace https://datamarket.azure.com/home.

Moor, D., Grubenmann, T., Seuken, S., & Bernstein, A. (2015). A double auction for querying the web of data. In *The Third Conference on Auctions, Market Mechanisms and Their Applications*.

Myerson, R. B. (1981). Optimal auction design. *Math. Oper. Res.*, *6*(1), 58--73.

Myerson, R. B., & Satterthwaite, M. A. (1983). Efficient mechanisms for bilateral trading. *Journal of Economic Theory*, *29*(2), 265--281.

Nisan, N., Roughgarten, T., Tardos, E., & Vazirani, V. (2007). *Algorithmic Game Theory*. Cambridge University Press.

Ramel (2016). Microsoft closing azure datamarket. Application Development Trends Magazine, https://adtmag.com/articles/2016/11/18/azure-datamarket-shutdown.aspx.

Schomm, F., Stahl, F., & Vossen, G. (2013). Marketplaces for data: An initial survey. *ACM SIGMOD*, *42*(1), 15--26.

Thomson-Reuters (2015). Thomson reuters annual report. Tech. rep..

Tirole, J., & Laffont, J.-J. (1993). *A Theory of Incentives in Procurement and Regulation*. MIT Press.

Varian, H. R. (1995). Pricing information goods. Tech. rep., University of Michigan.

Varian, H. R. (1997). Versioning information goods. Tech. rep., University of California.

W3C (2014). Linked Data. https://www.w3.org/standards/semanticweb/data.

5 Data Markets with Dynamic Arrival of Buyers and Sellers

> The only reason for time is so that everything doesn't happen at once.
>
> Albert Einstein

The content of this chapter will soon (in slightly revised form) appear in:

> Moor, D. (2019). Data Markets with Dynamic Arrival of Buyers and Sellers. In *Proceedings of the 14th Workshop on Economics of Networks, Systems and Computation*, NetEcon, Phoenix, AZ.

Data Markets with Dynamic Arrival of Buyers and Sellers

Dmitry Moor
University of Zurich
Zurich, Switzerland
dmoor@ifi.uzh.ch

ABSTRACT

We propose a market design solution for a market for distributed data. The main challenges addressed by our solution are (1) different data providers produce different databases that can be joined to produce answers for users' queries; (2) data providers have high fixed costs for producing their databases; and (3) buyers and sellers can arrive dynamically to the market. Our design relies on using a Markov chain with states corresponding to different numbers of allocated databases. The transition probabilities between different states are governed by the payments suggested by the market platform to the data providers. The main challenge in this setting is to guarantee *dynamic incentive compatibility*, i.e., to ensure that buyers and sellers are not incentivized to arrive late to the market or to misreport their costs or values. To achieve this, we disentangle the payments suggested by the market platform to the sellers from the posted prices exposed to the buyers. We prove that the buyer-optimal payments that are exposed to sellers are non-increasing which prevents late arrivals of sellers. Further, we demonstrate that the posted prices exposed to buyers constitute a martingale process (i.e., late arrivals lead to the same expected price). Finally, we show that our design guarantees zero expected average budget deficit and we perform a number of simulations to validate our model.

CCS CONCEPTS

• **Applied computing** → **Economics**; *Marketing*;

KEYWORDS

Market Design, Data Markets, Dynamic Markets

ACM Reference Format:
Dmitry Moor. 2018. Data Markets with Dynamic Arrival of Buyers and Sellers. In *Proceedings of NetEcon 2019: The 14th Workshop on the Economics of Networks, Systems and Computation (NetEcon 2019).* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Many datasets on the Web are unstructured. This means that they can be easily interpreted by humans but not by machines. Imposing some structure on the data by publishing it as a database and linking it to other databases can help machines to make sense of the content of the data. This significantly reduces the effort of humans for the search, analysis and making predictions based on this data by delegating many of these tasks to the machine. This naturally results in great benefits for society (see Bernstein et al. [2016]).

The technology for producing and querying such structured and distributed data already exists and used in numerous areas (e.g., [W3C 2014]). Despite of all its potential benefits, this technology is not highly utilized. One of the reasons for that is the lack of financial incentives of data providers to publish their data in a structured format. This happens because the high fixed costs that the data providers incur for producing their databases, structuring the data and linking it to the datasets of other data providers can never be recouped ([Moor et al. 2019]). As a result, a different system of incentives is required to compensate the data providers. In this paper, we propose such a system by designing a market for distributed data.

1.1 Call for Data Markets

In recent years, there were numerous attempts to design a market for data. Koutris et al. [2015] aim at designing a market for selling different views of a database while satisfying a *no-arbitrage* constraint. However, their approach does not easily extend to domains when users join data produced by multiple data providers.

Moor et al. [2015, 2019] and Agarwal et al. [2019] emphasize the importance of joining data coming from different data providers. They argue that the combinatorial preferences of buyers is a crucial feature for data markets as many databases can complement each other. As a result, the buyer who can access more databases gets a more precise and thus, valuable answer for his query.

However, none of these studies consider the dynamics of the data market. While in many combinatorial markets the dynamics may not play a critical role, data markets are inherently dynamic.[1] This means that both buyers and sellers in these markets arrive regularly and can strategically delay their arrivals if they expect to be better off by doing so. Due to the combinatorial nature of preferences of buyers such delays can have a dramatic effect on the operation of the market. Indeed, the late arrivals of sellers may result in a very low surplus reached by the buyers who arrive earlier and thus, can access only very few databases. In our work, we focus on both of these aspects, i.e., on the complementary nature of the data and on the dynamics of the market.

1.2 Overview of our approach

In this paper, we propose a model for a dynamic data market. We focus on the following challenges: (1) the data providers have high

[1] For example, combinatorial spectrum auctions typically happen once in several years (Cramton [2013]). Within this time frame the technology can change dramatically making it impractical for the bidders to misreport their bids based on the expected outcome of one of the future auctions.

NetEcon 2019, June 28, 2019, Phoenix, AZ

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/1122445.1122456

fixed costs for producing their databases; (2) the databases can be complementary for the buyers, i.e., joining two databases generates some additional value for the buyers; (3) buyers and sellers can arrive to the market over time and can strategically decide when to arrive.

We adopt a similar approach as proposed by Moor et al. [2019], i.e., we design a market that aims at optimizing the surplus of buyers while guaranteeing that the sellers' costs for producing their databases are compensated. In contrast to [Moor et al. 2019], we design a market that uses posted prices for both sellers and buyers. The rationale for this design decision is twofold. On the one hand, the market with posted prices has a very simple interface for possibly non-sophisticated buyers and sellers. On the other hand, the restriction of using the posted prices results in a much simpler strategic behavior of buyers and sellers. Indeed, in this case, they do not have to compute their optimal bid but can simply respond to the proposed posted prices.[2]

The market platform in our market plays the role of a *regulator*, i.e., it decides on which sellers to allocate, which queries to execute and how much the buyers need to pay to the sellers. Thus, on the one hand, the market platform computes payments for the sellers and allows the sellers to respond to these payments. If the seller's cost is smaller than the proposed payment, then the seller gets allocated, i.e., she creates and delivers her database to the market platform. On the other hand, the market platform computes the posted prices (per query) that are exposed to the buyers. The market platform then receives the buyers' queries, executes them, collects the respective amounts of money from the buyers and transfers them to the sellers.

We demonstrate how to compute the payments suggested to the sellers and the posted prices for the buyers in a way that neither sellers nor buyers have an incentive to strategically delay their arrival or misreport their costs or values. This guarantees *dynamic incentive compatibility*. Furthermore, we argue that while the traditional notion of *budget balancedness* is incompatible with dynamic incentive compatibility, our market design still satisfies *zero expected average budget deficit*. In other words, we show that, as the number of databases grows, the expected budget deficit per seller decreases to zero. Finally, we validate our approach via simulations.

2 FORMAL MODEL

We assume that time is discrete and we consider an infinite horizon problem where $t = 0, 1, 2, \ldots, \infty$ are the consecutive time steps. We let $N \in \mathbb{N}$ be the maximum number of databases that the market platform can allocate.

Sellers. Data providers arrive to the market independently at different time steps. At every time step at most one data provider with the new database can arrive with probability r.[3] Each data provider can produce a single database.

We let $\theta_i = \langle a_i, c_i \rangle$ be the type of the data provider i. Here, $a_i \in \mathbb{N}$ is the arrival time of the data provider, i.e., the time step when the data provider obtains her data; c_i is the fixed cost that the data provider incurs for producing the database out of her data. We assume that all c_i are drawn independently from the cumulative distribution $F(c)$, $f(c)$ is the corresponding density function. We assume that c_i includes mainly the *labor* cost for producing the database, i.e., costs for setting up the database, structuring the data, linking the data against other existing databases etc.[4] We assume that θ_i is a private knowledge of the data provider and let $\hat{\theta}_i = \langle \hat{a}_i, \hat{c}_i \rangle$ be the reported type of the data provider.

Consider the data provider i who obtains her data at time t, i.e., $a_i = t$. This data provider can decide to structure her data and to produce a database. We assume that the database can be produced immediately after the data provider gets her data. As the data provider is strategic, she can decide to deliver her database to the market at a different *reported arrival time* $\hat{a}_i \neq a_i$ if she expects to be better off by doing so. We impose the following assumption on early arrivals:

Assumption 1 (Sellers' Limited Misreports). *For every data provider i it must hold $\hat{a}_i \geq a_i$.*

This assumption is not too restrictive as data providers cannot produce and deliver their databases before they obtain the actual data (which happens at time t).

Let $X_t \in \mathbb{N}$ denote the number of databases allocated at time t, $X_0 = 0$. Also, let $p(X_t)$ be the payment that the market platform is willing to pay for the new database when $X_t - 1$ databases have already been allocated. Then,

$$X_{t+1} = X_t + \sum_{\hat{\theta}_i \colon \hat{a}_i = t+1} \mathbb{1}\{\hat{c}_i \leq p(X_t + 1)\}.$$

Informally, this means that a new database is allocated at time $t + 1$ if there is an arrival of a new data provider at time $t + 1$ and the cost of the data provider is not larger than the payment $p(X_t + 1)$ proposed by the market platform.

We assume that data providers have quasi-linear utility functions, i.e., the present value of the utility of the data provider i who obtains her data at time a_i but decides to deliver it at time $\hat{a}_i$ is $u_i(\theta_i, \hat{\theta}_i) = -c_i + \delta^{\hat{a}_i - a_i} p(X_{\hat{a}_i})$; here $p(X_{\hat{a}_i})$ is the payment paid by the market platform to the data provider; $\delta \in (0, 1)$ is the constant discount rate for money. [5]

Buyers. Generally speaking, at every time step multiple buyers with different queries can arrive. Each buyer is willing to pay a certain amount of money for an answer for his query. To keep our model simple, instead of considering the demand of each buyer separately, we consider an aggregate demand of all buyers. In other words, we assume that at every time step there is a single risk-neutral *aggregate* buyer willing to get an answer for his question by submitting a query. In what follows, we will always refer to the aggregate buyer as simply a "buyer".

A buyer who arrives with his question at time t can strategically submit his query late at time $\hat{t} \neq t$ if he expects to be better off by doing so. We assume that the buyer cannot submit his query before he gets his question to ask:

[2] In what follows we will use the word (posted) *payments* for sellers and *posted prices* for buyers.

[3] In practice, this can be achieved by making time intervals small enough. Considering a continuous time model with a Poisson arrival process is a possible future extension.

[4] We assume zero marginal costs, i.e., the electricity costs, the costs of maintaining the data etc.

[5] Notice, that the sellers discount only their future payments but not the costs. This follows from the fact that these are the "labor" costs and must be indexed over time with the same rate δ (i.e., c_i is the constant present value of the future labor costs).

Assumption 2 (Buyers' Limited Misreports). *Buyers cannot arrive earlier, i.e., $\hat{t} \geq t$.*

In this setting, the instantaneous utility of the buyer who gets his question at time t but submits his query at time $\hat{t}$ is $\mathcal{U}_t(\hat{t}) = \gamma^{\hat{t}-t}\big(V(X_{\hat{t}}) - \tau_{\hat{t}}(X_{\hat{t}})\big)$, where $\gamma \in [0, \delta)$ is the discount factor for the buyer's utility;[6] $\tau_{\hat{t}}(X_{\hat{t}})$ is the posted price faced by the buyer at time $\hat{t}$ if $X_{\hat{t}}$ databases are allocated. Observe that in our setting, the posted prices $\tau_t(X_t)$ depend on the number of allocated databases and thus, constitute a stochastic process (see Section 4 for more details). The expected value $V(.)$ of the buyer for the answer for his query depends on the number of allocated databases X_t. We assume that $V(.)$ is concave and strictly increasing. This reflects the fact that the larger is the number of available databases, the more informative (and thus, valuable) an answer for the buyer's query can be. Furthermore, the marginal value of an additional database becomes smaller as the number of allocated databases grows. Thus, such a shape of $V(.)$ captures the complementarity aspect of the buyers' preferences and the diminishing value of additional databases. Important here is that all databases are assumed to be *homogeneous*, i.e., they have similar values for possibly different groups of individual buyers. This assumption excludes the "junk" data, i.e., the data that has no value for any individual buyer. We elaborate on this value model in Appendix B. We also assume that $V(.)$ is known by the market platform.[7]

Remark 1. *In practice, the value of each individual (not aggregate) buyer for his query can depend not only on the number of allocated databases X_t but also on the identities of those databases. While these preferences of individual buyers may be very diverse (and generally unknown), the aggregate preferences are typically much simpler to predict. This idea was discussed by [Bakos and Brynjolfsson 1999] who suggested bundling of information goods as a way to obtain consumers' valuations for those goods. With this interpretation, in our model the buyers pay for an access to a bundle of databases. Under mild assumptions one can let the value of the buyers for such an access to be concave and strictly increasing in the number of databases in the bundle.*

Market Platform. Our design relies on modeling the dynamics of the market via a Markov chain. The states of this Markov chain correspond to different numbers of allocated databases. The transition probabilities are defined by the arrival rates of the sellers and the payments suggested by the market platform to the sellers. To compute these payments, we adopt a similar approach as proposed by [Moor et al. 2019], i.e., we aim at optimizing the total expected future discounted utility of buyers while guaranteeing that the fixed costs of the allocated sellers are compensated. The rationale for such a market design objective comes from the fact that in data markets, the sellers can be "monopolists" for their data. Thus, the market platform should play the role of a *regulator* that prevents the rent extracting behavior of the sellers (see [Moor et al. 2019]).

[6] Buyers are typically not willing to wait for a long time before getting their queries answered. Consequently, γ is normally much smaller than δ. We also assume that γ is a common knowledge.

[7] Similarly to [Moor et al. 2019], the buyers' side of the market is thick and one can easily sample buyers to learn their valuations. In practice, such learning can be performed by the market platform by iteratively updating its belief about $V(.)$ when observing the responses of the buyers for the posted prices. The design of the respective learning procedure, however, is outside the scope of this paper.

Formally, we can think about our market platform as a Markov chain with $N+1$ states. A state is characterized by the number of databases being allocated at this state. Assume that at time t the market platform is in the state $X_t \in \{0, 1, ..., N\}$. At this state, the market platform announces a payment $p(X_t + 1)$ for the seller arriving next. Data providers observe the proposed payments and decide whether to produce their databases. We set explicitly $p(N + 1) = 0$ to indicate that in the terminal state, no further databases can be allocated.

We impose a number of constraints on our market design.

Definition 1 (DIC for Sellers). *The mechanism is **dynamic incentive compatible for sellers** if for any seller i and $\forall \theta_i, \hat{\theta}_i$ that satisfy Assumption 1 we have $u_i(\theta_i, \theta_i) \geq \mathbb{E}_{X_{\hat{a}_i}}[u_i(\theta_i, \hat{\theta}_i)|X_{a_i}]$.*

In words, we say that the mechanism is dynamic incentive compatible for sellers, if neither seller can expect to get a higher utility at any of the future states $X_{\hat{a}_i}$ by misreporting her cost or by delaying her arrival.

Definition 2 (DIC for Buyers). *The mechanism is **dynamic incentive compatible for buyers** if $\exists t^* > 0$ s.t., for any $t \geq t^*$ we have $\mathcal{U}_t(t) \geq \mathbf{E}_{X_{\hat{t}}}[\mathcal{U}_t(\hat{t})|X_t]$ for any $\hat{t}, t$ that satisfy Assumption 2.*

In words, we say that the mechanism is dynamic incentive compatible for buyers if once the market gets sufficiently large (i.e., many databases are available), the buyers cannot expect to get a higher utility by delaying their arrival. The latter definition rules out some corner cases that can occur when the market just starts operating, i.e., during the interval $[0, t^*]$ when only very few databases are available.

Given these design constraints, we can now formally define the transitions of the Markov chain. Let i, j be the states of the Markov chain and let $P = [P_{ij}]_{(N+1)\times(N+1)}$ be the stochastic transition matrix of this Markov chain with

$$P_{ij} = \begin{cases} rF\big(p(i+1)\big), & \texttt{if } j = i+1 \\ 1 - rF\big(p(i+1)\big), & \texttt{if } j = i \\ 0, & \texttt{otherwise.} \end{cases} \quad (1)$$

Thus, the Markov chain transitions from the state i to the state $i+1$ if there is an arrival of a seller (with probability r) and the cost of the seller is not larger than the payment proposed to this seller (which happens with probability $F\big(p(i+1)\big)$). Let $P^n = P \cdot P^{n-1}$, $n = 2, 3, ...$

A commonly used property of *budget balancedness* (see, e.g., [Mas-Colell et al. 1995]) can be informally stated as follows: A mechanism is budget balanced if the total amount of money paid to the sellers net the total amount collected from the buyers is equal to zero. Observe that in our setting, the notion of budget balancedness is not compatible with DIC for Buyers. Indeed, assume that at some time step the posted price for the buyer is $\tau_0 > 0$ and all N databases are already allocated. If at some time step $t^* > 0$ the mechanism is budget balanced, then for any $\epsilon > 0$ and for any $t \geq t^*$ the amount of money that should be collected from buyers is smaller than ϵ. Let us choose $\epsilon < \tau_0$. Then, the posted price at any time $t \geq t^*$ must be smaller than ϵ and consequently, smaller than τ_0. Thus, the buyer who does not discount the future strongly (i.e., $\gamma \approx 1$) would always prefer to wait until t^* to submit his query. This violates the DIC for Buyers.

Thus, instead of focusing on the traditional notion of budget balancedness, we aim at achieving *zero expected average budget deficit*, i.e., we show that the *shortfall* per seller decreases as the number of databases increases. We can define this property formally in the following way: Let $\tilde{p}(t)$ be the present value (at time t) of all the past payments that have been already paid to the allocated sellers up to time t. Similarly, let $\tilde{\tau}(t)$ be the present value of all the payments made by the buyers up to time t. Thus, the *budget deficit* at time t can be defined as $BD(t) = \tilde{p}(t) - \tilde{\tau}(t)$.

Definition 3 (Expected Budget Deficit). *The expected budget deficit is*

$$\mathbb{E}[BD] = \lim_{t\to\infty} \mathbb{E}_{\theta_i}[BD(t)],$$

where the expectation is over different types of sellers θ_i.

In words, the expected budget deficit is equal to the expected residual amount of money that even in the limit cannot be collected from the buyers to fully compensate the sellers. Zero average expected budget deficit requires that this loss per-seller becomes negligibly small as the market grows, i.e.,

Definition 4 (Zero Expected Av. Budget Deficit). *The mechanism has **zero average expected budget deficit** if*

$$\lim_{N\to\infty} \frac{\mathbb{E}[BD]}{N} = 0.$$

The mechanism we propose is individually rational for buyers (sellers) as they can always opt-out if the proposed price (payment) is larger (smaller) than their value (cost).

3 COMPUTING PAYMENTS TO SELLERS

Remember that the market platform maximizes the total expected future discounted surplus of buyers. Let v_k^* be the maximal expected total future discounted surplus of buyers when k databases are already allocated. Consider the Bellman equations for the market platform:

$$v_N^* = \frac{V(N)}{1-\gamma} \tag{2}$$

$$v_{k-1}^* = \max_{p(k)} \Big\{ V(k-1) + \gamma r F\big(p(k)\big)\big(v_k^* - p(k)\big) + \Big(1 - rF\big(p(k)\big)\Big)\gamma v_{k-1}^* \Big\} \tag{3}$$

for $k = 1, ..., N$. Informally, the maximal expected total future discounted surplus of buyers in the state $k-1$ of the Markov chain is equal to the immediate "reward" in this state, i.e., $V(k-1)$, plus the discounted expected future maximal surplus in the next state. The latter one depends on whether the Markov chain stays in the state $k-1$ (i.e., if no allocation happens) or if it transitions to the state k.

The first-order conditions imply

$$v_k^* - p^*(k) - v_{k-1}^* = \frac{F\big(p^*(k)\big)}{f\big(p^*(k)\big)}. \tag{4}$$

Now, we can rewrite

$$v_{k-1}^* = V(k-1) + r\gamma \frac{F^2(p^*(k))}{f(p^*(k))} + \gamma v_{k-1}^* \tag{5}$$

Equations (4) and (5) constitute a system of $2N$ non-linear equations with $2N$ unknowns.[8] The solution of these equations gives us the payments for sellers $p^*(k)$ at every state k of the Markov chains (along with the values v_k^*).

Now, we claim that the sellers have no incentive to arrive late. This follows from the fact that in such a setting the payments proposed by the market platform can only decrease with time. This proves dynamic incentive compatibility for sellers. The following theorem states this formally.

Definition 5. *We say that a distribution $f(c)$ is **strongly regular** if $\frac{F(c)}{f(c)}$ is monotone and strictly increasing.*

Theorem 1. *If $f(.)$ is strongly regular, then the mechanism is dynamic incentive compatible for sellers.*

Proof. We first show that if $f(.)$ is strongly regular, then $p(1), p(2), \ldots$ weakly decreases with time. The Bellman equations can be rewritten as follows:

$$v_{k-1}^* - v_{k-2}^* = \frac{1}{1-\gamma}\Big[V(k-1) - V(k-2) + \gamma r\Big(\frac{F^2\big(p^*(k)\big)}{f\big(p^*(k)\big)} - \frac{F^2\big(p^*(k-1)\big)}{f\big(p^*(k-1)\big)}\Big)\Big]$$

for $k = 1, ..., N$. Using Equation (4) we can rewrite:

$$p^*(k-1) + \frac{F\big(p^*(k-1)\big)}{f\big(p^*(k-1)\big)} + \frac{\gamma r}{1-\gamma}\frac{F^2\big(p^*(k-1)\big)}{f\big(p^*(k-1)\big)} = \frac{1}{1-\gamma}\Big[V(k-1) - V(k-2) + \gamma r \frac{F^2\big(p^*(k)\big)}{f\big(p^*(k)\big)}\Big].$$

By induction, we see that $p(N+1) = 0, p(N) > 0$; the l.h.s. is a strictly increasing function while the r.h.s. gets larger as $k \to 0$ (due to the concavity of $V(.)$ and the induction hypothesis $p^*(k) \geq p^*(k+1)$). Thus, the solution for $p^*(k-1)$ must also get larger as $k \to 0$, i.e., $p^*(k-1) \geq p^*(k)$.

Finally, for any $\theta_i, \hat{\theta}_i$ that satisfy Assumption 1 we have

$$u_i(\theta_i, \hat{\theta}_i) = -c_i + \delta^{\hat{a}_i - a_i} p(X_{\hat{a}_i}) \leq -c_i + \delta^{\hat{a}_i - a_i} p(X_{a_i}) \leq -c_i + p(X_{a_i}) = u_i(\theta_i, \theta_i).$$

Q.E.D. □

4 COMPUTING PRICES FOR BUYERS

Observe that if we set the posted prices for the buyer equal to the payments for sellers, i.e., $\tau_t(X_t) = p(X_t)$, then the mechanism cannot satisfy DIC for Buyers. Indeed, as we have shown in Theorem 1, the payments $p(X_t)$ can only decrease with time. In this case, the posted price would also only decrease. This would incentivize the buyers to arrive late which violates DIC for Buyers. Therefore, we need to disentangle the posted prices exposed to the buyer from the payments paid to the sellers. There are two main requirements to constructing such posted prices:

R1. The posted prices τ_t must guarantee DIC for Buyers;
R2. τ_t and $p(X_t)$ must satisfy zero expected average budget deficit.

[8] We solve it with the Newton method.

In this section, we show how to construct a pricing scheme satisfying these two requirements.

First, the solution of Equations (4) and (5) allows us to compute the transition matrix P as defined in Equation (1). Now, let $\tilde{\pi}(k)$ denote the present value of the *future total payment* to sellers when the Markov chain is in the state k. Thus, for each k we can compute the expected value of $\tilde{\pi}(k)$ at this state as follows:

$$\begin{aligned}\mathbb{E}[\tilde{\pi}(k)] =&\delta r F\big(p(k+1)\big)\big(p(k+1)+\mathbb{E}[\tilde{\pi}(k+1)]\big)+ \\ &\delta\Big(1-rF\big(p(k+1)\big)\Big)\mathbb{E}[\tilde{\pi}(k)],\quad \forall k=0,1,...,N.\end{aligned} \tag{6}$$

In words, if the Markov chain is in the state k, then two things can happen. Either an allocation happens, and therefore, the Markov chain transitions to the state $k+1$. In this case, the market platform must make an "immediate" payment $p(k+1)$ and expects to make a future payment of $\mathbb{E}[\tilde{\pi}(k+1)]$. Alternatively, no allocation happens. In this case, the market platform stays in the state k and expects to make a future payment of $\mathbb{E}[\tilde{\pi}(k)]$. Thus, the present value of the expected future total payment to the sellers in state k is equal to the discounted convex combination of the two aforementioned terms. Equations (6) constitute a system of $N+1$ linear equations with $N+1$ unknowns $\mathbb{E}[\tilde{\pi}(k)]$, $k=0,...,N$.

Now remember, that P^n_{ij} is the probability that the Markov chain transitions from the state i to the state j within n time intervals. Thus, to satisfy the requirement R1 at time $t=0$ we must have

$$\begin{aligned}&\tau_0(0)+\delta\big(P_{00}\tau_1(0)+P_{01}\tau_1(1)+...\big)+ \\ &\delta^2\big(P^2_{00}\tau_2(0)+P^2_{01}\tau_2(1)+P^2_{02}\tau_2(2)+...\big)+...=\mathbb{E}[\tilde{\pi}(0)].\end{aligned} \tag{7}$$

To satisfy the requirement R2 we compute the prices in a way that at any time step t and any allocation of databases X_t at time t, the expected future posted price at any possible future time interval is equal to the current posted price (i.e. to the posted price at time t). Thus, for $t=0$ we set $P_{00}\tau_1(0)+P_{01}\tau_1(1)+...=\tau_0(0)$, $(P^2_{00}\tau_2(0)+P^2_{01}\tau_2(1)+P^2_{02}\tau_2(2)+...)=\tau_0(0)$ etc. for any $\hat{t}>t$. Now, we can simplify the Equation (7): $\tau_0(0)=(1-\delta)\mathbb{E}[\tilde{\pi}(0)]$. Generally, if at time t the Markov chain is in the state k, we set

$$\tau_t(k)=(1-\delta)\Big(\mathbb{E}[\tilde{\pi}(k)]+\tilde{p}(t)-\frac{\tilde{\tau}(t-1)}{\delta}\Big). \tag{8}$$

Finally, the overall mechanism looks as follows:

Dynamic Data Market Mechanism

Payments to Sellers: At time $t=0$, solve Equations (4) and (5):

$$v^*_{k-1}=V(k-1)+r\gamma\frac{F^2(p^*(k))}{f(p^*(k))}+\gamma v^*_{k-1},$$

$$v^*_k-p^*(k)-v^*_{k-1}=\frac{F\big(p^*(k)\big)}{f\big(p^*(k)\big)},\qquad k=1,...,N.$$

Here, $p^*(k)$ is the payment proposed by the market platform for the k'th database, $k=1,...,N$.

Allocation of Sellers: At each time $t>0$, a seller with cost c_i may arrive and respond to $p^*(X_{t-1}+1)$. If $c_i\le p^*(X_{t-1}+1)$, the seller is allocated, $X_t=X_{t-1}+1$. Otherwise, the seller is not allocated, $X_t=X_{t-1}$.

Posted Prices for Buyers: At each time t, the market platform computes the posted price for this interval according to Equation (8):

$$\tau_t(X_t)=(1-\delta)\Big(\mathbb{E}[\tilde{\pi}(X_t)]+\tilde{p}(t)-\frac{\tilde{\tau}(t-1)}{\delta}\Big).$$

The dynamic incentive compatibility for buyers follows from the following theorem.

Theorem 2. *The process $\tau_t(X_t)$ is a martingale.*

Proof. See Appendix A. □

To complete the proof of DIC for Buyers, observe that $\gamma V(X_{t+1})\le\gamma V(X_t+1)=\gamma(V(X_t+1)-V(X_t))+\gamma V(X_t)$. From concavity of $V(.)$ it follows that as X_t gets sufficiently large, the difference $(V(X_t+1)-V(X_t))$ gets small. This fact together with Theorem 2 proves that if the market is large enough, the buyers do not get more value from delaying their arrival. Formally,

$$\begin{aligned}\mathbf{E}[\mathcal{U}_t(\hat{t}=t+1)|X_t]=&\mathbf{E}[\gamma V(X_{t+1})|X_t]-\gamma\mathbf{E}[\tau_{\hat{t}}(X_{\hat{t}})|X_t]\le \\ &\underbrace{\gamma\mathbb{E}[V(X_t+1)-V(X_t)]}_{\texttt{Goes to 0 as } X_t \texttt{ grows.}}+\gamma\mathcal{U}_t(t).\end{aligned}$$

Finally, the following theorem shows that the proposed mechanism satisfies zero expected average budget deficit.

Theorem 3. *The dynamic data market mechanism has zero expected average budget deficit.*

Proof. See Appendix A. □

5 EXPERIMENTS

To validate our model, we carry out a number of simulations of the proposed market under different simulation scenarios. We assume that for all scenarios, the costs of sellers are drawn from the uniform distribution, $c_i\sim U[0,1]$. We further assume that the value of the buyer is $V(X_t)=\sqrt{X_t}$. The discount rate is $\delta=0.9$.

Payments for Sellers. First, we perform a simulation with $N=100$ databases while varying r and γ. Figure 1 illustrates the payment $p(i)$ for the newly arriving database $i\le N$ when $(i-1)$ databases are already allocated. Here, we vary $r\in\{0.1,0.5,1.0\}$ while fixing $\gamma=0.9$. In line with our results proved in Theorem 1, the payments decrease over time. From this figure, we also see that as the arrival rate r gets smaller, the market platform suggests higher payments to the sellers. This result follows form the fact that as the probability of arrival of a seller decreases, the opportunity cost of the market platform for waiting increases. Indeed, if at time t the seller does not arrive, and $X_t=X_{t-1}$, then the buyer enjoys a smaller value of $V(X_{t-1})$ instead of the value $V(X_{t-1}+1)$ he could have enjoyed if the seller arrived at time t and delivered her database. Thus, the market platform "loses" the possible higher value of the buyer and consequently, has a higher opportunity cost for waiting. Due to the increased opportunity costs, the market platform increases the payments.

Now, let us look into the dependency of the payments to the sellers on the discount factor γ of the buyer. Figure 2 illustrates the payments of the market platform for the i'th database when $(i-1)$ databases are already allocated. Here, we fix $r=1$ and vary $\gamma\in\{0.1,0.5,0.9\}$. The figure demonstrates that the stronger the buyer discounts the future, the smaller the payments proposed to

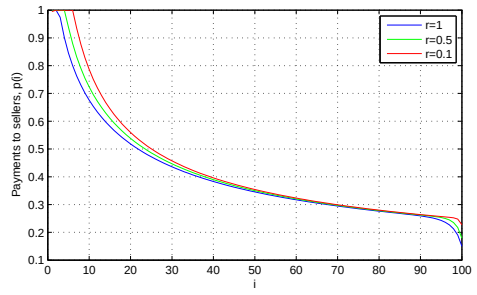

Figure 1: Payment for the i'th allocated database for different arrival rates r of sellers.

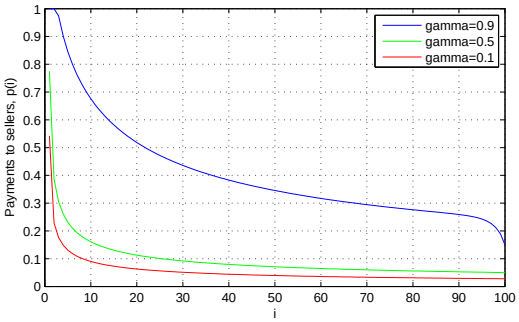

Figure 2: Payment for the i'th allocated database for different discount factors γ of buyers.

the sellers by the market platform. The explanation of this phenomenon comes from a similar opportunity cost argument: Indeed, stronger discounting of the future value of the buyer decreases the opportunity cost of "losing" the future buyer's surplus. Thus, the opportunity cost of not allocating the seller now gets smaller. Consequently, the payments suggested by the market platform to the sellers must also decrease.

Posted Prices for the Buyer. We illustrate the posted prices exposed to the buyer by generating 10 trajectories corresponding to the process $\tau_t(X_t)$. To achieve this, we sample 10 different arrival scenarios and costs c_i. We set $\gamma = 0.5$, $r = 1$, $N = 100$. We then let the simulated sellers arrive to the market and respond to the suggested payments. At each time step t we compute the number of allocated databases X_t as well as the posted price $\tau_t(X_t)$ according to Equation (8). Figure 3 (top) illustrates the different trajectories corresponding to the martingale process of the posted price $\tau_t(X_t)$ while Figure 3 (bottom) demonstrates the respective trajectories of the process X_t. From comparing the Figure 3 (top) with the Figure 3 (bottom) we see that if an allocation does not happen at time t (i.e., the trajectory of X_t has a plateau), then the posted price $\tau_t(X_t)$ decreases. If an allocation happens at time t, then there is a respective spike in the posted price.

Expected Average Budget Deficit. We illustrate the convergence of the expected average budget deficit, $\frac{\mathbb{E}[BD]}{N}$, to zero as the number of allocated databases N grows. Figure 4 illustrates our findings. Here, we sample 1000 different trajectories corresponding to different arrivals and costs of sellers. We then compute the mean values and the standard errors of the resulting expected average budget deficit. As expected, the result goes in hand with our Theorem 3.

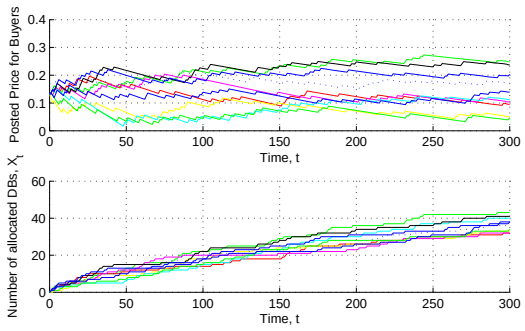

Figure 3: Trajectories of the posted price $\tau_t(X_t)$ (top) and the number of allocated databases X_t (bottom). For every trajectory X_t, the respective trajectory $\tau_t(X_t)$ is depicted with the same color.

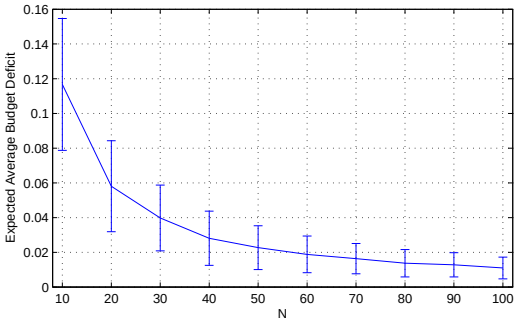

Figure 4: Expected Average Budget Deficit for different numbers of N. Here, $r = 1$, $\gamma = 0.5$.

6 CONCLUSIONS

In this paper, we have studied the dynamics of the combinatorial data market. We proposed a mechanism that optimizes the expected future discounted surplus of buyers while compensating the fixed costs of allocated sellers and satisfying the two key properties: dynamic incentive compatibility and zero expected average budget deficit. We further studied the proposed mechanism in a simulation environment. Our results confirm our intuition regarding the changes in prices and in the budget deficit when slightly changing the parameters of the mechanism. In future work, we are planning to expand these simulations and to study a number of further economic properties of the proposed mechanism.

REFERENCES

Anish Agarwal, Munther Dahleh, and Tuhin Sarkar. A marketplace for data: An algorithmic solution. Technical report, Massachusetts Institute of Technology, Working Paper, 2019.

Yannis Bakos and Erik Brynjolfsson. Bundling information goods: Pricing, profits, and efficiency. *Management Science*, 45(12):1613–1630, 1999. URL http://EconPapers.repec.org/RePEc:inm:ormnsc:v:45:y:1999:i:12:p:1613-1630.

Dirk Bergemann, Alessandro Bonatti, and Alex Smolin. The design and price of information. *American Economic Review*, 108(1):1–48, January 2018. doi: 10.1257/aer.20161079. URL http://www.aeaweb.org/articles?id=10.1257/aer.20161079.

Abraham Bernstein, James Hendler, and Natalya Noy. A new look at the semantic web. *Commun. ACM*, 59(9):35–37, August 2016. ISSN 0001-0782. doi: 10.1145/2890489. URL http://doi.acm.org/10.1145/2890489.

P. Cramton. Spectrum auction design. *Review of Industrial Organization*, 42(2):030–190, March 2013.

Paraschos Koutris, Prasang Upadhyaya, Magdalena Balazinska, Bill Howe, and Dan Suciu. Query-based data pricing. In *Journal of the ACM (JACM)*, volume 62, October

2015.
Andreu Mas-Colell, Michael D. Whinston, and Jerry R. Green. *Microeconomic Theory.* Oxford University Press, New York, 1995.
Dmitry Moor, Tobias Grubenmann, Sven Seuken, and Abraham Bernstein. A double auction for querying the web of data. In *The Third Conference on Auctions, Market Mechanisms and Their Applications*, 2015.
Dmitry Moor, Sven Seuken, Tobias Grubenmann, and Abraham Bernstein. The design of a combinatorial data market. Technical report, Faculty of Business, Economics and Informatics, University of Zurich, Working Paper, 2019.
W3C. Linked Data. https://www.w3.org/standards/semanticweb/data, 2014.

A PROOFS

Lemma 1. *For any state $k = 0, \ldots, N$ the following inequality holds:*

$$\mathbb{E}[\tilde{\pi}(k)] - \mathbb{E}[\tilde{\pi}(k+1)] \leq p(k+1).$$

Proof. Follows from Equation (6). We can rewrite

$$\Big(1 + \frac{1-\delta}{\delta r F\big(p(k+1)\big)}\Big)\mathbb{E}[\tilde{\pi}(k)] - \mathbb{E}[\tilde{\pi}(k+1)] = p(k+1). \quad (9)$$

Here, $1 + \frac{1-\delta}{\delta r F\left(p(k+1)\right)} \geq 1$. Therefore,

$$\mathbb{E}[\tilde{\pi}(k)] - \mathbb{E}[\tilde{\pi}(k+1)] \leq p(k+1).$$

□

Lemma 2. *For any time $t > 0$ it holds $\tilde{p}(t) - \tilde{\tau}(t) > 0$.*

Proof. We proceed by induction. For $t = 1$, the payment $p(1)$ is maximal and the result holds, i.e., $BD(1) = p(1) - \frac{\tau_0(0)}{\delta} - \tau_1(\ell) > 0$. Consider an arbitrary time $t > 1$ and $X_t = \ell$. We have

$$\begin{aligned}
BD(t) =& \tilde{p}(t) - \tilde{\tau}(t) = \tilde{p}(t) - \frac{\tilde{\tau}(t-1)}{\delta} - \\
& (1-\delta)\Big(\mathbb{E}[\tilde{\pi}(\ell)] + \tilde{p}(t) - \frac{\tilde{\tau}(t-1)}{\delta}\Big) = \\
& \delta\tilde{p}(t) - \tilde{\tau}(t-1) - (1-\delta)\mathbb{E}[\tilde{\pi}(\ell)] = \\
& \delta\Big(\tilde{p}(t) - \frac{\tau(t-1)}{\delta}\Big) - (1-\delta)\mathbb{E}[\tilde{\pi}(\ell)] = \\
& \delta\Big(\frac{\tilde{p}(t-1)}{\delta} + p(X_t) - \frac{\tilde{\tau}(t-1)}{\delta}\Big) - (1-\delta)\mathbb{E}[\tilde{\pi}(\ell)] = \\
& BD(t-1) + \delta p(X_t) - (1-\delta)\mathbb{E}[\tilde{\pi}(\ell)].
\end{aligned}$$

Now, consider two cases: $p(X_t) = p(\ell)$ and $p(X_t) = p(\ell+1)$ (i.e., dependent on whether there is an allocation has happened at time t).

In the former case, using Equation (9) we can rewrite:

$$\begin{aligned}
BD(t) =& BD(t-1) + \Big(\delta + \frac{1-\delta}{rF\big(p(\ell)\big)}\Big)\mathbb{E}[\tilde{\pi}(\ell-1)] - \\
& \delta\mathbb{E}[\tilde{\pi}(\ell)] - (1-\delta)\mathbb{E}[\tilde{\pi}(\ell)] = \\
& BD(t-1) + \Big(\delta + \frac{1-\delta}{rF\big(p(\ell)\big)}\Big)\mathbb{E}[\tilde{\pi}(\ell-1)] - \mathbb{E}[\tilde{\pi}(\ell)] \\
& \geq BD(t-1) \geq 0.
\end{aligned}$$

In the latter case,

$$\begin{aligned}
BD(t) =& BD(t-1) + \Big(\delta + \frac{1-\delta}{rF\big(p(\ell)\big)}\Big)\mathbb{E}[\tilde{\pi}(\ell)] - \\
& \delta\mathbb{E}[\tilde{\pi}(\ell+1)] - (1-\delta)\mathbb{E}[\tilde{\pi}(\ell)] = \\
& BD(t-1) + \Big(\delta + \frac{1-\delta}{rF\big(p(\ell)\big)} - (1-\delta)\Big)\mathbb{E}[\tilde{\pi}(\ell)] - \\
& \delta\mathbb{E}[\tilde{\pi}(\ell+1)] \geq BD(t-1) \geq 0.
\end{aligned}$$

Q.E.D. □

Theorem 3 1. *The mechanism has zero expected average budget deficit.*

Proof. We first prove the following lemma:

Lemma 4. *For any state $k = 0, \ldots, N$ the following inequality holds:*

$$\mathbb{E}[\tilde{\pi}(k)] - \mathbb{E}[\tilde{\pi}(k+1)] \leq p(k+1). \quad (10)$$

Proof. See Appendix A. □

Now, we can show that for any time $t > 0$, the budget deficit is non-negative.

Lemma 5. *For any time $t > 0$ it holds $\tilde{p}(t) - \tilde{\tau}(t) > 0$.*

Proof. See Appendix A. □

Now, let us consider the budget deficit at time t, i.e., $BD(t) = \tilde{p}(t) - \tilde{\tau}(t)$. We know that $BD(0) = 0 - \tilde{\tau}(0) = -(1-\delta)\mathbb{E}[\tilde{\pi}(0)]$. The expected budget deficit at time $t = 1$ is

$$\begin{aligned}
\mathbb{E}[BD(1)] =& rF\big(p(1)\big)\Big(p(1) - \tau_1(1)\Big) - \frac{\tau_0(0)}{\delta} - \Big(1 - rF\big(p(1)\big)\Big)\tau_1(0) = \\
& rF\big(p(1)\big)\Big(p(1) - \tau_1(1) + \tau_1(0)\Big) - \frac{\tau_0(0)}{\delta} - \tau_1(0).
\end{aligned}$$

Observe, that $\tau_1(1) = \tau_1(0) + (1-\delta)(p(1) + \mathbb{E}[\tilde{\pi}(1)] - \mathbb{E}[\tilde{\pi}(0)])$. Thus, we can rewrite

$$\begin{aligned}
\mathbb{E}[BD(1)] =& rF\big(p(1)\big)\Big(p(1) - (1-\delta)(p(1) + \mathbb{E}[\tilde{\pi}(1)] - \mathbb{E}[\tilde{\pi}(0)])\Big) - \\
& \frac{\tau_0(0)}{\delta} - \tau_1(0) = \\
& rF\big(p(1)\big)\Big(\mathbb{E}[\tilde{\pi}(0)] - \mathbb{E}[\tilde{\pi}(1)]\Big) + \delta rF\big(p(1)\big)\Big(p(1) + \mathbb{E}[\tilde{\pi}(1)]\Big) - \\
& \delta rF\big(p(1)\big)\mathbb{E}[\tilde{\pi}(0)] - \frac{\tau_0(0)}{\delta} - \tau_1(0) = \\
& rF\big(p(1)\big)\Big(\mathbb{E}[\tilde{\pi}(0)] - \mathbb{E}[\tilde{\pi}(1)]\Big) + \delta rF\big(p(1)\big)\Big(p(1) + \mathbb{E}[\tilde{\pi}(1)]\Big) + \\
& \delta\Big(1 - rF\big(p(1)\big)\Big)\mathbb{E}[\tilde{\pi}(0)] \\
& -\delta\mathbb{E}[\tilde{\pi}(0)] - \frac{\tau_0(0)}{\delta} - (1-\delta)\Big(\mathbb{E}[\tilde{\pi}(0)] - \frac{\tau_0(0)}{\delta}\Big) \\
& = rF\big(p(1)\big)\Big(\mathbb{E}[\tilde{\pi}(0)] - \mathbb{E}[\tilde{\pi}(1)]\Big) + BD(0).
\end{aligned}$$

From Lemma 4 it follows that

$$\mathbb{E}[BD(1)] \leq BD(0) + rF\big(p(1)\big)p(1).$$

Now, consider the expected budget deficit at time $t > 1$:

$$\begin{aligned}
\mathbb{E}[BD(t)] =& \mathbb{E}[BD(t-1)] + \\
& \sum_{\ell} \Pr(state = \ell)\Big[rF\big(p(\ell+1)\big)\big(p(\ell+1) - \tau_t(\ell+1)\big) + \\
& \Big(1 - rF\big(p(\ell+1)\big)\Big)\big(-\tau_t(\ell)\big)\Big].
\end{aligned}$$

Observe, that

$$\tau_t(\ell+1) = \tau_t(\ell) + (1-\delta)\Big(\mathbb{E}[\tilde{\pi}(\ell+1)] - \mathbb{E}[\tilde{\pi}(\ell)] + p(\ell+1)\Big).$$

Thus, we can rewrite

$$
\begin{aligned}
&rF\big(p(\ell+1)\big)\Big(p(\ell+1)-\tau_t(\ell+1)\Big)+\\
&\Big(1-rF\big(p(\ell+1)\big)\Big)(-\tau_t(\ell))=\\
&rF\big(p(\ell+1)\big)\Big(p(\ell+1)-\tau_t(\ell+1)+\tau_t(\ell)\Big)-\tau_t(\ell)=\\
&rF\big(p(\ell+1)\big)\Big(\delta p(\ell+1)-(1-\delta)(\mathbb{E}[\tilde{\pi}(\ell+1)]-\mathbb{E}[\tilde{\pi}(\ell)])\Big)\\
&\qquad-\tau_t(\ell)=\\
&rF\big(p(\ell+1)\big)\Big(\mathbb{E}[\tilde{\pi}(\ell)]-\mathbb{E}[\tilde{\pi}(\ell+1)]\Big)+\\
&\delta rF\big(p(\ell+1)\big)\Big(p(\ell+1)+\mathbb{E}[\tilde{\pi}(\ell+1)]\Big)+\\
&\delta\Big(1-rF\big(p(\ell+1)\big)\Big)\mathbb{E}[\tilde{\pi}(\ell)]-\delta\mathbb{E}[\tilde{\pi}(\ell)]-\tau_t(\ell)=\\
&rF\big(p(\ell+1)\big)\Big(\mathbb{E}[\tilde{\pi}(\ell)]-\mathbb{E}[\tilde{\pi}(\ell+1)]\Big)+\\
&\qquad\mathbb{E}[\tilde{\pi}(\ell)](1-\delta)-\tau_t(\ell)\leq\\
&rF\big(p(\ell+1)\big)\Big(\mathbb{E}[\tilde{\pi}(\ell)]-\mathbb{E}[\tilde{\pi}(\ell+1)]\Big)-\frac{1-\delta}{\delta}BD(t-1)<\\
&rF\big(p(\ell+1)\big)\Big(\mathbb{E}[\tilde{\pi}(\ell)]-\mathbb{E}[\tilde{\pi}(\ell+1)]\Big).
\end{aligned}
$$

Here, the last inequality follows directly from Lemma 5. Now, we can rewrite

$$
\begin{aligned}
&\mathbb{E}[BD(t)]<\mathbb{E}[BD(t-1)]+\\
&\qquad\max_{\ell}\Big\{rF\big(p(\ell+1)\big)\Big(\mathbb{E}[\tilde{\pi}(\ell)]-\mathbb{E}[\tilde{\pi}(\ell+1)]\Big)\Big\}\leq\\
&\qquad\mathbb{E}[BD(t-1)]+\max_{\ell}\Big\{rF\big(p(\ell+1)\big)p(\ell+1)\Big\}.
\end{aligned}
$$

Which implies that the expected budget deficit grows slower than linearly. Thus,

$$\lim_{N\to\infty}\frac{\mathbb{E}[BD(t)]}{N}=0.$$

Q.E.D. □

Theorem 2 1. *The process $\tau_1(X_1), \tau_2(X_2), \ldots$ is a martingale.*

Proof. Let $X_t = s$. We want to show that $\mathbb{E}[\tau_{t+1}(\ell)|s] = \tau_t(s)$. Precisely,

$$
\begin{aligned}
&\mathbb{E}[\tau_{t+1}(\ell)|s]=rF\big(p(s+1)\big)\tau_{t+1}(s+1)+\\
&\qquad\Big(1-rF\big(p(s+1)\big)\Big)\tau_{t+1}(s)=\\
&\qquad(1-\delta)\Big[rF\big(p(s+1)\big)\Big(\mathbb{E}[\tilde{\pi}(s+1)]+\frac{\tilde{p}(t)}{\delta}+p(s+1)-\tilde{\tau}(t)\Big)\\
&\qquad\quad+\Big(1-rF\big(p(s+1)\big)\Big)\Big(\mathbb{E}[\tilde{\pi}(s)]+\frac{\tilde{p}(t)}{\delta}-\tilde{\tau}(t)\Big)\Big]=\\
&\qquad(1-\delta)\Big[rF\big(p(s+1)\big)\Big(\mathbb{E}[\tilde{\pi}(s+1)]+p(s+1)\Big)+\\
&\qquad\quad\Big(1-rF\big(p(s+1)\big)\Big)\mathbb{E}[\tilde{\pi}(s)]+\frac{\tilde{p}(t)}{\delta}-\tilde{\tau}(t)\Big]=\\
&\qquad(1-\delta)\Big[\frac{\mathbb{E}[\tilde{\pi}(s)]}{\delta}+\frac{\tilde{p}(t)}{\delta}-\tilde{\tau}(t)\Big]=\\
&\qquad\frac{1-\delta}{\delta}\Big[\mathbb{E}[\tilde{\pi}(s)]+\tilde{p}(t)-\delta\tilde{\tau}(t-1)-\tau_t(s)\Big]=\\
&\qquad\frac{1-\delta}{\delta}\Big[\frac{\tau_t(s)}{1-\delta}-\tau_t(s)\Big]=\tau_t(s).
\end{aligned}
$$

Q.E.D. □

B VALUE MODEL

We assume that buyers acquire the data to make certain predictions about the state of the world. If the prediction is good, they get a high reward, $R_H \in \mathbb{R}^+$. Otherwise, they get a low reward $R_L \in \mathbb{R}^+$, $R_L < R_H$ (this is similar to the model of [Bergemann et al. 2018]). Due to the inherent uncertainty about the world, a buyer without any additional information faces a lottery A in which he can get the high reward with $\Pr(R_H) = \tilde{P}_A$ or the low reward with $\Pr(R_L) = 1 - \tilde{P}_A$. Thus, the expected reward of the risk-neutral buyer who does not acquire any data is

$$R_A = \tilde{P}_A R_H + (1 - \tilde{P}_A) R_L.$$

We assume that better predictions of the state of the world lead to higher chances of getting the high reward R_H for the buyer, i.e., to a different lottery with a higher expected reward. Thus, in order to improve his prediction, the buyer can purchase the data. As typically the number of different databases N is much larger than the number of databases relevant for answering each buyer's query, we assume that every buyer is willing to join at most two databases.

In what follows, we show that as the number of available databases N increases, the aggregate value of buyers for a yet another database also increases. Generally speaking, allocating an additional complementary database may result in either concave or convex aggregate value function $V(.)$. However, we demonstrate that for larger numbers of N, the aggregate value function $V(.)$ is more likely to be concave even for strongly complementary databases. This follows from the fact, that as N grows, the buyers are still joining only a very small number of databases comparing to N (two in our example). Therefore, there is no exponential blow up in the values of buyers for answers for their queries *on average*. We show this more formally by demonstrating a number of examples for $N \in \{0, 1, 2, 3\}$. Development of a more general formal value model is the future extension of this work.

Single database. Assume that accessing a single database results in a perfect prediction of the state of the world with probability $\tilde{P}_1$. Conversely, the database is useless with probability $(1 - \tilde{P}_1)$.[9] This induces a new lottery B, such that the expected reward of this new lottery is

$$R_B = (1 - \tilde{P}_1)R_A + \tilde{P}_1 R_H = R_A + \tilde{P}_1(1 - P_A)(R_H - R_L).$$

Consequently, the expected willingness to pay for the single database is

$$R_B - R_A = \tilde{P}_1(R_H - R_A) = \tilde{P}_1(1 - \tilde{P}_A)(R_H - R_L). \quad (11)$$

Two databases. Now, assume that there are two databases available for the buyers. As before, let $\tilde{P}_1$ be the probability to make a perfect prediction regarding the state of the world by addressing *either* of the two databases. Implicit here is the *homogeneity* assumption, i.e., the assumption that this probability is the same for both databases. This obviously excludes the case of the "junk" data, i.e., the data that is not valuable by neither buyer.

Furthermore, let $\tilde{P}_2$ be the probability to make a perfect prediction by joining the two databases.[10] Let C be the respective lottery faced by buyers who can access both databases. The expected reward of this new lottery is

$$R_C = R_H\Big(2\tilde{P}_1 - \tilde{P}_1^2 + \tilde{P}_2\Big) + R_A\Big(1 - (2\tilde{P}_1 - \tilde{P}_1^2 + \tilde{P}_2)\Big).$$

Here, the total probability of getting the high reward by accessing any single database is $2\tilde{P}_1 - \tilde{P}_1^2$ while the probability of getting the high reward by joining the two databases is $\tilde{P}_2$. In this case, the expected willingness to pay for the second database is

$$R_C - R_B = \Big(\tilde{P}_1(1 - \tilde{P}_1) + \tilde{P}_2\Big)(1 - \tilde{P}_A)(R_H - R_L). \quad (12)$$

Comparing this with Equation (11) we see that as long as $\tilde{P}_2 \leq \tilde{P}_1^2$, we have $R_C - R_B \leq R_B - R_A$. This means that as long as the two databases complement each other (i.e., $\tilde{P}_2 > 0$) but are not "strongly" complementary, the expected willingness of the buyers to pay for the second database is smaller than the one for the first database. This implies concavity of the value function $V(.)$ of the buyer.

Contrary, if $\tilde{P}_2 > \tilde{P}_1^2$, the expected willingness of the buyers to pay for the second database is larger than their expected willingness to pay for the first one. This means that if the two databases are strongly complementary, the resulting value function $V(.)$ can become convex. However, as we show below, this effect decreases as we increase N.

Three databases. Consider the case when there are three databases available. Remember, that buyers are willing to join at most two of them. In this case, the probability to get the high reward is $(3\tilde{P}_1 - 3\tilde{P}_1^2 + \tilde{P}_1^3) + (3\tilde{P}_2 - 3\tilde{P}_2^2 + \tilde{P}_2^3)$. Here, the first term reflects the probability of getting the high reward by accessing a single database. The second term reflects the probability of receiving the high reward when joining two different databases.[11] In this case, the willingness to pay for the third database is

$$R_D - R_C = \Big(\tilde{P}_1(1 - \tilde{P}_1)^2 + 2\tilde{P}_2 - 3\tilde{P}_2^2 + \tilde{P}_2^3\Big)(1 - \tilde{P}_A)(R_H - R_L).$$

Comparing this equation with Equation (12) we see that in this case, to achieve concavity of $V(.)$ we only need to show that

$$\tilde{P}_1(1 - \tilde{P}_1)^2 + 2\tilde{P}_2 - 3\tilde{P}_2^2 + \tilde{P}_2^3 \leq \tilde{P}_1(1 - \tilde{P}_1) + \tilde{P}_2.$$

This latter constraint is trivially less strict than the constraint $\tilde{P}_2 \leq \tilde{P}_1^2$ obtained for the case of two databases. Thus, as the number of available databases N grows and the buyers are willing to join only a small subset of these databases, the expected willingness of the buyers to pay for a yet another database decreases even for databases with stronger complementarity properties. Thus, given the assumption of homogeneity of databases and the restricting the number of databases that can be joined by buyers, the aggregate value function $V(.)$ can be assumed to be concave for large N.

[9] Another way of thinking about this is that the fraction of the population of buyers $\tilde{P}_1$ can get an answer for their questions using this single database.
[10] If the two databases are complementary, it must be that $\tilde{P}_2 > 0$.
[11] Here we consider only the symmetric case, when different ways of joining the data may lead to obtaining the perfect prediction regarding the state of the world.

Curriculum Vitae

Personal Information

Name	Dmitrii Moor
Date of Birth	April 1, 1990
Place of Birth	Moscow, USSR
Nationality	Russian

Professional Experience

07/2018 -- 10/2018	*Summer Intern*, IBM Research
10/2013 -- 09/2014	*Research Assistant*, ETH Zurich
03/2011 -- 10/2013	*Software Engineer*, IBM

Education

2014 -- 2019	*PhD Candidate*
	Computation and Economics Research Group
	University of Zurich
2017 -- 2018	*Graduate Diploma in Economics*
	University of London
2010 -- 2012	*MSc in Computer Science*
	Bauman Moscow State Technical University
2006 -- 2010	*BSc in Computer Science*
	Bauman Moscow State Technical University